VISIT…

LANZAROTE
Caliente.COM

САНКТ–ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ

КАФЕДРА ВЫЧИСЛИТЕЛЬНОЙ ФИЗИКИ

В.А.БУСЛОВ , С.Л.ЯКОВЛЕВ

ЧИСЛЕННЫЕ МЕТОДЫ II. РЕШЕНИЕ УРАВНЕНИЙ

КУРС ЛЕКЦИЙ

САНКТ–ПЕТЕРБУРГ

2001

Утверждено на заседании кафедры

вычислительной физики

печатается по решению методической комиссии

физического факультета СПбГУ

А В Т О Р Ы : В.А.БУСЛОВ, С.Л.ЯКОВЛЕВ

Р Е Ц Е Н З Е Н Т : докт. физ.-мат. наук С.Ю.СЛАВЯНОВ

Настоящее издание является второй частью курса лекций по численным методам, читавшихся на протяжении ряда лет авторами в первом семестре II курса физического факультета СПбГУ.

В пособии принята нумерация формул по главам. Приведенная библиография частично представляет собой источник справочного материала, но, в основном, рассчитана на дальнейшее изучение численных методов.

Глава 1

Системы уравнений

1.1 Решение нелинейных уравнений

Задачу нахождения решений уравнений можно формулировать различными способами. Например, как задачу на нахождение корней: $f(x) = 0$, или как задачу на нахождение неподвижной точки: $F(x) = x$. При этом, в зависимости от формулировки задачи, удобно применять те или иные способы решения. Рассмотрим сначала одномерную ситуацию.

1.1.1 Одномерный случай

Метод деления пополам

Простейшим методом нахождения корней уравнения $f(x) = 0$ является метод деления пополам или *дихотомия*. Предположим, мы нашли две точки x_0 и x_1, такие что $f(x_0)$ и $f(x_1)$ имеют разные знаки, тогда между этими точками, если $f \in C^0$, находится хотя бы один корень функции f . Поделим отрезок $[x_0, x_1]$ пополам и введем точку $x_2 = \frac{x_0+x_1}{2}$. Либо $f(x_2)f(x_0) \leq 0$, либо $f(x_2)f(_1) \leq 0$. Оставим ту половину отрезка, для которой значения на концах имеют разные знаки. Теперь этот отрезок делим пополам и оставляем ту его часть, на границах которой функция имеет разные знаки, и так далее, до достижения требуемой точности.

К достоинствам метода деления пополам следует отнести его высокую надежность и простоту, при этом от функции требуется только непрерывность. Порядок сходимости метода линейный, на каждом шаге точность возрастает вдвое.

Недостатком метода является тот факт, что прежде чем начать его применение, необходимо предварительно найти две точки, значения функции в которых имеют разные знаки. Очевидно, что метод неприменим для корней четной кратности. Он также не может быть обобщен на случай комплексных корней и на системы уравнений.

Метод простых итераций

Пусть $F : [a, b] \to [a, b]$ и F — сжатие: $|F(x) - F(y)| \leq q|x - y|$, $q < 1$ (в частности, тот факт, что F — сжатие, как легко видеть, означает, что $F \in C_{[a,b]}$). По теореме Банаха существует и единственна

неподвижная точка x^*. Она может быть найдена как предел простой итерационной процедуры

$$x^* = \lim_{n\to\infty} x_n \ , \quad x_{n+1} = F(x_n) \ ,$$

где начальное приближение x_0 — произвольная точка промежутка $[a,b]$. Если функция F дифференцируема, то удобным критерием сжатия является число $q = \sup\limits_{x\in[a,b]} |F'(x)| = ||F'||_C < 1$. Действительно, по теореме Лагранжа

$$|F(x) - F(y)| = |F'(\xi)||x-y| \le ||F'||_C|x-y| = q|x-y| \ .$$

Таким образом, если производная меньше единицы, то F является сжатием.

Условие $F([a,b]) \subseteq [a,b]$ существенно, ибо если, например, $F(x) \equiv 2$ на [0,1] , то неподвижная точка отсутствует, хотя производная равна нулю. Скорость сходимости зависит от величины q. Чем меньше q, тем быстрее сходимость.

Пример. Решить уравнение: $x^2 = a$. Здесь, если в качестве F взять функцию $F(x) = \frac{a}{x}$, то соответствующая итерационная процедура будет иметь вид: $x_{n+1} = \frac{a}{x_n}$. Как нетрудно убедиться, метод итераций в данном случае расходится при любой начальной точке x_0, не совпадающей с собственно неподвижной точкой $x^* = \sqrt{a}$. Однако можно в качестве F предложить и более хитрую функцию с той же неподвижной точкой. Пусть $F(x) = \frac{1}{2}[x + \frac{a}{x}]$. Соответствующая итерационная процедура здесь имеет вид: $x_{n+1} = \frac{1}{2}[x_n + \frac{a}{x_n}]$. Эти итерации сходятся к неподвижной точке для любого начального приближения $x_0 \in (0,\infty)$:

$$\begin{cases} F(x) = \frac{a}{x} \ , & x_{n+1} = \frac{a}{x_n} \ , \\ F(x) = \frac{1}{2}[x + \frac{a}{x}] \ , & x_{n+1} = \frac{1}{2}[x_n + \frac{a}{x_n}] \ , \quad x_n \to x^* \ . \end{cases}$$

Действительно, в первом случае $F'(x_n) = -\frac{a}{x_n^2}$, т.е. для выполнения условия $F'(x_n) < 1$ необходимо чтобы $x_n^2 > a$, но тогда $|F'(x_{n+1})| = |\frac{a}{x_{n+1}^2}| = \frac{a}{\frac{a^2}{x_n^2}} = \frac{x_n^2}{a} > 1$. Таким образом, отображение $F(x) = \frac{a}{x}$ сжатием не является.

Для $F(x) = \frac{1}{2}[x + \frac{a}{x}]$, где неподвижная точка та же самая, ситуация другая. Здесь, хотя формально производная может быть довольно большой (при малых x), однако уже на следующем шаге она будет меньше 1. Убедимся в этом:

$$F'(x_{n+1}) = \frac{1}{2}\left[1 - \frac{a}{x_{n+1}^2}\right] = \frac{1}{2}\left[1 - \frac{a}{\frac{1}{2}(x_n + \frac{a}{x_n})^2}\right] =$$

$$= \frac{1}{2}\left[1 - \frac{\frac{a}{x_n^2}}{\frac{1}{2}(1 + \frac{a}{x_n^2})^2}\right] = \frac{1}{2}\frac{(1 + \frac{a}{x_n^2})^2 - \frac{2a}{x_n^2}}{(1 + \frac{a}{x_n^2})^2} = \frac{1}{2}\frac{1 + \left(\frac{a}{x_n^2}\right)^2}{(1 + \frac{a}{x_n^2})^2} < \frac{1}{2} \ ,$$

т.е. такой итерационный процесс всегда сходится.

1.1.2 Метод Ньютона

Метод Ньютона или *касательных* заключается в том, что если x_j — некоторое приближение к корню x_* уравнения $f(x) = 0$, $f \in C^1$, то следующее приближение определяется как корень касательной к функции $f(x)$, проведенной в точке x_j. Таким образом, в уравнении касательной $f'(x_j) = \frac{y - f(x_j)}{x - x_j}$ необходимо положить $y = 0$ и $x = x_{j+1}$, то есть

$$x_{j+1} = x_j - \frac{f(x_j)}{f'(x_j)} \ .$$

Поскольку метод Ньютона представляет собой метод простых итераций при $F(x) = x - \frac{f(x)}{f'(x)}$, то нетрудно убедиться, что при $f \in C^2$ существует окрестность корня, в которой $|F'| < 1$. Действительно,

$$F' = 1 - \frac{(f')^2 - ff''}{(f')^2} = \frac{ff''}{(f')^2} ,$$

то если x_* корень кратности α , то в его окрестности $f(x) \approx a(x-x_*)^\alpha$ и, следовательно, $F'(x_*) = \frac{\alpha-1}{\alpha}$. Заметим, что если x_* — простой корень, то сходимость метода касательных квадратичная (то есть порядок сходимости равен 2). Убедимся в этом. Поскольку $x_{j+1} - x_* = x_j - x_* - \frac{f(x_j)}{f'(x_j)}$, то

$$\frac{x_{j+1} - x_*}{(x_j - x_*)^2} = \frac{1}{x_j - x_*} - \frac{f(x_j)}{f'(x_j)(x_j - x_*)^2} = \frac{1}{x_j - x_*} -$$

$$- \frac{f'(x_*)(x_j - x_*) + \frac{1}{2} f''(x_*)(x_j - x_*)^2 + o\left((x_j - x_*)^2\right)}{(x_j - x_*)^2 \left[f'(x_*) + f''(x_*)(x_j - x_*) + o(x_j - x_*)\right]} ,$$

откуда

$$\lim_{j \to \infty} \frac{x_{j+1} - x_*}{(x_j - x_*)^2} = \frac{f''(x_*)}{2f'(x_*)} .$$

Таким образом, сходимость метода Ньютона очень быстрая. При этом без всяких изменений метод обобщается на комплексный случай. Если корень x_* является корнем второй кратности и выше, то, как нетрудно убедиться, порядок сходимости сразу падает и становится линейным.

К недостаткам метода Ньютона следует отнести его локальность, поскольку он гарантированно сходится при произвольном стартовом приближении только, если везде выполнено $|ff''|/(f'^2) < 1$, в противной ситуации сходимость есть лишь в некоторой окрестности корня. Другим недостатком метода Ньютона является тот факт, что на каждом шаге необходимо заново вычислять производную.

1.1.3 Метод секущих

Чтобы избежать вычисления производной, метод Ньютона можно упростить, заменив производную на разностную, вычисленную по двум предыдущим итерациям, что эквивалентно замене функции $f(x)$ на интерполяционный полином, проходящий через точки x_j и x_{j-1} . При этом, итерационный процесс принимает вид

$$x_{j+1} = x_j - \frac{f_j(x_j - x_{j-1})}{f_j - f_{j-1}} ,$$

где $f_j = f(x_j)$. Это двухшаговый итерационный процесс, поскольку использует для нахождения последующего приближения два предыдущих. Порядок сходимости метода секущих, естественно, ниже, чем у метода касательных и равен в случае однократного корня $d = \frac{\sqrt{5}+1}{2}$. Убедимся в этом, считая для удобства, что $x_* = 0$.

$$\frac{f_j(x_j - x_{j-1})}{f_j - f_{j-1}} = \frac{[f'_* x_j + \frac{1}{2} f''_* x_j^2 + O(x_j^3)](x_j - x_{j-1})}{f'_*(x_j - x_{j-1}) + \frac{1}{2} f''_*(x_j^2 - x_{j-1}^2) + O(x_j^3 - x_{j-1}^3)} =$$

$$= x_j \left[\frac{1 + \frac{f''_*}{2f'_*} x_j + O(x_j^2)}{1 + \frac{f''_*}{2f'_*}(x_j + x_{j-1}) + O(x_j^2)}\right] = x_j \left[1 - \frac{f''_*}{2f'_*} x_{j-1} + O(x_j^2)\right] .$$

Таким образом, с точностью до бесконечно малых более высокого порядка

$$x_{j+1} = x_j - \frac{f_j(x_j - x_{j-1})}{f_j - f_{j-1}} = \frac{f''_*}{2f'_*} x_j x_{j-1} + O(x_j^3) .$$

Отбрасывая остаточный член, получаем рекуррентное соотношение $x_{j+1} = \alpha x_j x_{j-1}$, $\alpha = \frac{f''_*}{2f'_*}$, решение которого естественно искать в виде $x_{j+1} = \alpha^c x_j^d$. После подстановки имеем: $cd = 1$ и $d^2 - d - 1 = 0$, откуда в силу того, что для сходимости необходимо, чтобы d было положительным, заключаем, что $d = \frac{\sqrt{5}+1}{2}$.

Поскольку знание производной не требуется, то при том же объёме вычислений в методе секущих (несмотря на меньший порядок сходимости) можно добиться большей точности, чем в методе касательных. Отметим, что вблизи корня приходится делить на малое число, и это приводит к потере точности (особенно в случае кратных корней), поэтому, выбрав относительно малое δ, выполняют вычисления до выполнения $|x_{j+1} - x_j| < \delta$ и продолжают их пока модуль разности соседних приближений убывает. Как только начнется рост, вычисления прекращают и последнюю итерацию не используют. Метод секущих становится неприменимым. Такая процедура определения момента окончания итераций называется приемом *Гарвика.*

Метод парабол

Рассмотрим трехшаговый метод, в котором приближение x_{j+1} определяется по трем предыдущим точкам x_j , x_{j-1} и x_{j-2} . Для этого заменим, аналогично методу секущих, функцию $f(x)$ интерполяционной параболой проходящей через точки x_j , x_{j-1} и x_{j-2} . В форме Ньютона она имеет вид

$$p_2(x) = f_j + f_{j-1,j}(x - x_j) + f_{j-2,j-1,j}(x - x_j)(x - x_{j-1}) \ .$$

Точка x_{j+1} определяется как тот из корней этого полинома, который ближе по модулю к точке x_j . Порядок сходимости метода парабол выше, чем у метода секущих, но ниже, чем у метода Ньютона. Важным отличием от ранее рассмотренных методов, является то обстоятельство, что даже если $f(x)$ вещественна при вещественных x и стартовые приближения выбраны вещественными, метод парабол может привести к комплексному корню исходной задачи. Этот метод очень удобен для поиска корней многочленов высокой степени.

Поиск всех корней

Общим недостатком почти всех итерационных методов нахождения корней является то, что они при однократном применении позволяют найти лишь один корень функции, при том — неизвестно какой. Чтобы найти другие корни, можно было бы брать новые стартовые точки и применять метод заново, но нет никакой гарантии, что при этом итерации сойдутся к новому корню, а не к уже найденному (если вообще сойдутся, как, скажем, возможно в методе Ньютона).

Для поиска других корней используется метод удаления корней. Пусть x_1 — корень функции $f(x)$, рассмотрим функцию $f_1(x) = \frac{f(x)}{x - x_1}$. Точка x_1 будет являться корнем функции $f_1(x)$ на единицу меньшей кратности, чем $f(x)$, при этом все остальные корни у функций $f(x)$ и $f_1(x)$ совпадают с учетом кратности. Применяя тот или иной метод нахождения корней к функции $f_1(x)$, мы найдем новый корень x_2 (который может в случае кратных корней и совпадать с x_1). Далее можно рассмотреть функцию $f_2(x) = \frac{f_1(x)}{x - x_2} = \frac{f(x)}{(x - x_1)(x - x_2)}$ и искать корни у неё. Повторяя указанную процедуру, можно найти все корни $f(x)$ с учетом кратности.

Заметим, что когда мы производим деление на тот или иной корень x_* , то в действительности мы делим лишь на найденное приближение x'_* , и, тем самым, несколько сдвигаем корни вспомогательной функции относительно истинных корней функции $f(x)$. Это может привести к значительным погрешностям, если процедура отделения применялась уже достаточное число раз. Чтобы избежать этого, с помощью вспомогательных функций вычисляются лишь первые итерации, а окончательные проводятся по исходной функции

$f(x)$, используя в качестве стартового приближения, последнюю итерацию, полученную по вспомогательной функции.

1.1.4 Многомерный случай

Метод простых итераций

Метод простых итераций (последовательных приближений) легко обобщается на случай системы нелинейных уравнений

$$f_k(x_1, x_2, \ldots, x_N) = 0 \ , \quad k = 1, 2, \ldots, N \ ,$$

или в векторной форме

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \ .$$

Эту систему удобно, как и в одномерном случае, записать в виде задачи на неподвижную точку

$$\mathbf{F}(\mathbf{x}) = \mathbf{x} \ .$$

Замечание. Нахождение такой формы записи может оказаться само по себе серьезной задачей. Необходимо добиться и того, чтобы отображение $\mathbf{F}$ являлось сжатием (для сходимости итераций) и, при этом, было эквивалентно исходной постановке.

Выбрав стартовое приближение, организуем итерации

$$\mathbf{x}^{(j+1)} = \mathbf{F}(\mathbf{x}^{(j)}) \ .$$

Если итерации сходятся, то они сходятся к одному из решений системы уравнений. Порядок сходимости простых итераций линейный. Действительно, пусть $\mathbf{x}^*$ решение, к которому сходятся итерации, тогда для каждой k-ой его компоненты

$$x_k^{(j+1)} - x_k^* = F_k(\mathbf{x}^{(j)}) - F_k(\mathbf{x}^*) = \sum_{l=1}^{N} \left[\frac{\partial F_k(\mathbf{z}^j)}{\partial x_l}\right] (x_l^{(j)} - x_l^*) \ ,$$

где $\mathbf{z}^j$ — некоторый вектор в направлении $\mathbf{x}^{(j)} - \mathbf{x}^*$, лежащий между этими точками. Отображение $\mathbf{F}$ будет являться сжатием, если норма матрицы производных (согласованная с нормой вектора в данном пространстве) $\left\{\frac{\partial F_k(\xi^k)}{\partial x_l}\right\}$ меньше единицы. Поскольку в конечномерном пространстве все нормы эквивалентны (а значит и последовательность сходящаяся по одной норме, будет сходиться и по любой другой), то достаточно это условие проверить для любой из норм матрицы с элементами $M_{kl} = \max |\frac{\partial F_k}{\partial x_l}|$, мажорирующей соответствующие нормы $\left\{\frac{\partial F_k(\xi^k)}{\partial x_l}\right\}$.

Улучшить сходимость метода последовательных приближений можно (хотя она по прежнему останется линейной), если уже найденные компоненты $x_k^{(j+1)}$ использовать для нахождения компонент этого же приближения $\mathbf{x}^{(j+1)}$ с номерами большими k , то есть организовать итерации следующим образом

$$x_{k+1}^{(j+1)} = F_k(x_1^{(j+1)}, x_2^{(j+1)}, \ldots, x_k^{(j+1)}, x_{k+1}^{(j)}, x_{k+2}^{(j)}, \ldots, x_N^{(j)}) \ .$$

Метод Ньютона

Метод Ньютона, являясь частным случаем метода простых итераций с вектор-функцией $\mathbf{F}$, равной

$$\mathbf{F}(\mathbf{x}) = \mathbf{x} - \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right]^{-1} \mathbf{f}(\mathbf{x}) \ ,$$

естественно обобщается на многомерный случай. Итерации по методу Ньютона имеют вид

$$\mathbf{x}^{(j+1)} = \mathbf{x}^{(j)} - \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right]^{-1}_{\mathbf{x}=\mathbf{x}^{(j)}} \mathbf{f}(\mathbf{x}^{(j)}) \ .$$

Проверка условий сходимости (то есть того, что норма матрицы производных $\partial \mathbf{F}/\partial \mathbf{x}$ меньше единицы) почти никогда не производится, поскольку требует большого объема вычислений. Сам же метод Ньютона обычно используют в несколько другой записи. Именно:

$$\left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{x}^{(j)}} \Delta\mathbf{x}^{(j)} = -\mathbf{f}(\mathbf{x}^{(j)}) \ , \quad \Delta\mathbf{x}^{(j)} = \mathbf{x}^{(j+1)} - \mathbf{x}^{(j)} \ .$$

Определяя из этой линейной системы (скажем методом Гаусса) вектор $\Delta\mathbf{x}^{(j)}$ и, соответственно, приближение $\mathbf{x}^{(j+1)}$, заново рассчитывают матрицу производных и продолжают итерации. Если начальное приближение выбрано удачно, то обычно достаточно всего нескольких итераций, поскольку сходимость квадратичная.

Методы спуска

Введём функцию $\Phi = \sum\limits_{j=1}^{N} |f_j(\mathbf{x})|^2$. Она ограничена снизу нулем и достигает своего глобального минимума (нуля) только в тех точках, где $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Таким образом, задача на поиск корней вектор-функции сводится к задаче на поиск минимума скалярной функции многих переменных, методы решения которой рассматриваются в соответствующей главе. Здесь лишь отметим, что эти методы называются методами спуска и являются сходящимися для гладких функций, однако точность их невелика, поэтому их естественно использовать для нахождения начального приближения с последующим использованием метода Ньютона. Важно также иметь в виду, что методы спуска могут сходится не к глобальному минимуму, а к одному из локальных (в зависимости от выбора стартовой точки), не отвечающих, разумеется, корням исходной задачи.

1.2 Решение линейных систем

1.2.1 Обусловленность линейных систем, погрешность

При решении абстрактной задачи $Ax = b$, где A — оператор произвольной природы, важным моментом является корректность ее постановки. Задача считается корректной, если решение существует и единственно и, кроме того, решение непрерывно зависит от входных данных (то есть, при $\Delta b \to 0$, Δx также стремится к нулю).

Однако и непрерывная зависимость от входных данных может иметь свои нюансы. Чем меньшее (большее) изменение решения вызывает вариация входных данных, тем более хорошо (плохо) *обусловленной* считается задача. Понятие обусловленности является тем более существенным для численных методов, поскольку на практике входные данные известны, как правило, с некоторой погрешностью. Кроме того, существуют ошибки округления, возникающие при вычислениях. Таким образом, формально корректная задача, являясь плохо обусловленной, может оказаться разрешимой столь неточно, что в этом будет отсутствовать практический смысл.

Чем можно охарактеризовать количественно обусловленность для линейных систем? Пусть A — квадратная $N \times N$-матрица. Рассмотрим задачу

$$A\mathbf{x} = \mathbf{b} \ .$$

Пусть также $|| * ||$ — какая-нибудь норма в $\mathbf{R}^N$ (например $||\mathbf{x}||=\max\limits_i |x_i|$, $= \sum |x_i|$, $= \sqrt{\sum x_i^2}$). Норма оператора A определяется стандартно

$$M = ||A|| = \max_{\mathbf{x} \neq \mathbf{0}} \frac{||A\mathbf{x}||}{||\mathbf{x}||} \ .$$

Обозначим $\mathbf{y} = A\mathbf{x}$ и введем число m по правилу

$$m = \min_{\mathbf{x} \neq \mathbf{0}} \frac{||A\mathbf{x}||}{||\mathbf{x}||} = \min_{\mathbf{y} \neq \mathbf{0}} \frac{||\mathbf{y}||}{||A^{-1}\mathbf{y}||} = \left(\max_{\mathbf{y} \neq \mathbf{0}} \frac{||A^{-1}\mathbf{y}||}{||\mathbf{y}||} \right)^{-1} = ||A^{-1}||^{-1} \ .$$

Величина $C(A) = \frac{M}{m} = ||A|| \cdot ||A^{-1}||$ называется *числом обусловленности*. Очевидно

1) $C(A) \geq 1$;

2) $C(\alpha A) = C(A)$;

3) если A — диагональная, то $C(A) = \frac{\max\limits_i |a_{ii}|}{\min\limits_i |a_{ii}|}$ (Для какой нормы, или для всех вышеприведенных?).

Чем меньше число обусловленности $C(A)$, тем лучше обусловлена система. Действительно, пусть $\Delta\mathbf{b}$ — вариация правой части, а $\Delta\mathbf{x}$ — соответствующее изменение решения. Тогда справедливо следующее неравенство

$$\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} \leq C(A) \frac{||\Delta\mathbf{b}||}{||\mathbf{b}||} \ .$$

Доказательство. Имеем: $A\mathbf{x} = \mathbf{b}$, $A(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$, $A\Delta\mathbf{x} = \Delta\mathbf{b}$. Так как

$$m \leq \frac{||A\Delta\mathbf{x}||}{||\Delta\mathbf{x}||} = \frac{||\Delta\mathbf{b}||}{||\Delta\mathbf{x}||} \ ,$$

то $||\Delta\mathbf{x}|| \leq \frac{1}{m}||\Delta\mathbf{b}||$. Аналогично, поскольку $A\mathbf{x} = \mathbf{b}$, то $||\mathbf{b}|| \leq M||\mathbf{x}||$. Объединяя два неравенства, окончательно получаем

$$\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} \leq \frac{M}{m} \frac{||\Delta\mathbf{b}||}{||\mathbf{b}||} \ .$$

1.2.2 Метод Гаусса

Одним из самых распространенных прямых методов решения систем линейных уравнений $A\mathbf{x} = \mathbf{b}$:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_N \end{pmatrix}$$

является метод Гаусса. Вначале исходная система приводится к верхнетреугольному виду. Это достигается следующей последовательностью преобразований (прямой ход метода Гаусса). Будем считать для удобства, что элементы a_{ij} исходной матрицы и компоненты вектора b_i есть, соответственно, элементы $a_{ij}^{(1)}$ первого шага преобразованной матрицы A_1 и преобразованного вектора $\mathbf{b}_1$: $A = A_1$, $\mathbf{b} = \mathbf{b}_1$. Далее, на втором шаге прибавим к второй строке первую, умноженную на $-\frac{a_{21}}{a_{11}} = c_{21}$. Аналогично поступим со всеми оставшимися строками, т.е. прибавим к каждой i-ой строке $i = 2, 3, \ldots, N$, первую, умноженную на коэффициент $_{i1} = -\frac{a_{i1}}{a_{11}}$. При этом соответственно изменится и вектор $\mathbf{b}_1$. Таким образом,

2 шаг) Имеем систему уравнений $A_2\mathbf{x} = \mathbf{b}_2$:

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2N}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & a_{N2}^{(2)} & \cdots & a_{NN}^{(2)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \cdots \\ b_N^{(2)} \end{pmatrix},$$

где $a_{ij}^{(2)} = a_{ij}^{(1)} + c_{i1}a_{1j}^{(1)}$, $b_i^{(2)} = b_i^{(1)} + c_{i1}b_1^{(1)}$, $i \geq 2$.

3 шаг) Прибавим к новой третьей строке новую вторую, умноженную на $c_{32} = -\frac{a_{32}^{(2)}}{a_{22}^{(2)}}$. То же самое сделаем с остальными строками $4, 5, \ldots, N$, т.е. прибавим к каждой i-ой строке вторую, умноженную на $c_{i2} = -\frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$, $i > 2$. При этом получим систему $A_3\mathbf{x} = \mathbf{b}_3$:

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & a_{N3}^{(3)} & \cdots & a_{NN}^{(3)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_N \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \cdots \\ b_N^{(3)} \end{pmatrix},$$

$(k+1)$-ый шаг) Здесь $a_{ij}^{(k+1)} = a_{ij}^{(k)} + c_{ik}a_{kj}^{(k)}$, $b_i^{(k+1)} = b_i^{(k)} + c_{ik}b_k^{(k)}$, где $c_{ik} = -\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$, $i, j > k$.

Поступая так и далее, на $(N-1)$-ом шаге получаем верхнетреугольную систему:

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & \cdots & a_{3N}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(4)} & \cdots & a_{4N}^{(4)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & a_{NN}^{(N)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \cdots \\ x_N \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ b_4^{(4)} \\ \cdots \\ b_N^{(N)} \end{pmatrix}.$$

При этом, мы также получили матрицу C переводных коэффициентов, имеющую вид:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ c_{21} & 0 & 0 & 0 & \cdots & 0 \\ c_{31} & c_{32} & 0 & 0 & \cdots & 0 \\ c_{41} & c_{42} & c_{43} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \ddots & \cdots \\ c_{N1} & c_{N2} & c_{N3} & \cdots & c_{NN-1} & 0 \end{pmatrix}.$$

Решение полученной треугольной системы $U\mathbf{x} = \mathbf{f}$ $(U = A_N$, $\mathbf{f} = \mathbf{b}_N)$, как легко видеть, имеет вид (обратный ход метода Гаусса)

$$x_N = \frac{f_{NN}}{U_{NN}}\ ,\ x_k = \frac{1}{U_{kk}}(f_k - \sum_{i=k+1}^{N} U_{ki}x_i)\ ,\ k = N, N-1, \ldots, 1\ .$$

Заметим, что при прямом ходе метода Гаусса может возникнуть ситуация, когда происходит деление на нуль, да и вообще, желательно не делить на малое число, чтобы не накапливалась ошибка. Поэтому метод Гаусса обычно проводят с *частичным выбором главного элемента*, то есть после каждого шага (пусть это

был k-й шаг) переставляют строки с номерами k, $k+1$, ..., N таким образом, чтобы на месте kk оказался элемент $a_{mk}^{(k)}$, наибольший из всех в k-ом столбце при $m > k$ (при этом, естественно, переставляются и компоненты вектора $\mathbf{b}$).

Можно для максимальной точности переставлять также и столбцы преобразуемой матрицы, чтобы на месте kk оказался максимальный элемент из всех с индексами больше, либо равными k. Эта процедура называется методом Гаусса с выбором главного элемента. Она несколько повышает точность по сравнению с частичным выбором главного элемента, но весьма неудобна, в том числе и для программирования, поскольку при перестановке строк компоненты искомого вектора $\mathbf{x}$ переставлять не надо, тогда как при перестановке столбцов надо переставлять и соответствующие компоненты вектора $\mathbf{x}$.

Опишем обратный ход метода Гаусса в несколько иной форме (треугольное разложение). Введем матрицы M_k по правилу

$$M_k = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ 0 & 0 & \cdots & c_{k+1,k} & \cdots & 0 \\ 0 & 0 & \cdots & c_{k+2,k} & \cdots & 0 \\ \dots & \dots & \dots & \dots & \ddots & \dots \\ 0 & 0 & \cdots & c_{N,k} & \cdots & 1 \end{pmatrix} .$$

На каждом шаге метода Гаусса получается некоторая промежуточная матрица $A_{k+1} = M_k M_{k-1} \dots M_1 A$, и вектор $\mathbf{f}_{k+1} = M_k M_{k-1} \dots M_1 \mathbf{b}$. Нетрудно видеть, что

$$U = \prod_{i=1}^{\overleftarrow{N-1}} M_i A \,, \quad \mathbf{f} = \prod_{i=1}^{\overleftarrow{N-1}} M_i \mathbf{b} \,; \quad U\mathbf{x} = \mathbf{f} \,, \; \det U = \prod_{i=1}^{N} U_{ii} = \det A \,.$$

Вопрос. Почему $\det U = \det A$?

Если производить также выбор главных элементов, то необходимо использовать оператор P перестановки индексов l и m, матричные элементы которого равны: $p_{ij} = 0$, $i, j \neq l, m$; $p_{im} = p_{mi} = 0$, $i \neq l$; $p_{li} = p_{il} = 0$, $i \neq m$; $p_{ml} = p_{lm} = 1$. При применении оператора перестановки индексов к матрице слева, меняются местами строки матрицы и компоненты свободного вектора ($PA\mathbf{x} = P\mathbf{b}$), если же его применить справа к матрице, то меняются местами ее столбцы и компоненты решения ($A \underbrace{PP}_{I} \mathbf{x} = \mathbf{b}$).

1.2.3 L-R разложение

Для решения задачи $A\mathbf{x} = \mathbf{b}$ несколько модифицируем ее. Именно, введем $N \times (N+1)$ матрицу

$$\mathbf{C} = \left(\begin{array}{c|c} & b_1 \\ \mathbf{A} & b_2 \\ & \vdots \\ & b_N \end{array} \right)$$

и вектор $\mathbf{X} = (x_1, x_2, \ldots, x_N, -1)^T$ размерности $(N+1)$, тогда исходная задача эквивалентна следующей

$$C\mathbf{X} = 0\ .$$

Представим C в виде $C = LR$, где L — нижнетреугольная $N \times N$ матрица

$$L = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{N1} & l_{N2} & \cdots & l_{NN} \end{pmatrix}\ ,$$

а R — $N \times (N+1)$-матрица:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1N} & r_{1,N+1} \\ 0 & 1 & r_{23} & \cdots & r_{2N} & r_{2,N+1} \\ 0 & 0 & 1 & \cdots & r_{3N} & r_{3,N+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & r_{N,N+1} \end{pmatrix}\ .$$

Как находить матрицы L и R?

1-й шаг) а) Умножим каждую строку матрицы L на первый столбец матрицы R, откуда $l_{i1} = c_{i1}$. Таким образом, мы определили первый столбец матрицы L.

б) Умножим первую строку L на каждый столбец R, откуда $r_{1i} = c_{1i}/l_{11}$, то есть определена первая строка R.

2-й шаг) а) Умножим каждую строку L (начиная со второй) на второй столбец R и определим второй столбец L: $l_{i2} = c_{i2} - l_{i1}r_{12}$.

б) Умножая вторую строку L на каждый столбец R, определяем вторую строку R: $r_{2i} = (c_{2i} - l_{21}r_{1i})/l_{22}$.

m-й шаг) Пусть известны первые $m-1$ столбец L и $m-1$ строка R, тогда при $i \geq m$

$$l_{im} = c_{im} - \sum_{k=1}^{m-1} l_{ik}r_{km}\ ,\quad r_{mi} = \frac{c_{mi} - \sum\limits_{k=1}^{m-1} l_{mk}r_{ki}}{l_{mm}}\ .$$

Теперь заметим, что вовсе нет необходимости решать задачу $C\mathbf{X} = 0$, а достаточно решить систему $R\mathbf{X} = 0$. Действительно, ранг матрицы R равен N, таким образом исходная матрица A и L вырождены или невырождены одновременно. Компоненты x_i находим последовательно, начиная с N-ой:

$$x_N = r_{N,N+1}\ ,\quad x_i = r_{i,N+1} - \sum_{k=i+1}^{N} r_{ik}x_k\ .$$

Вычисления по изложенному методу требуют в два раза меньший объем памяти, чем по методу Гаусса.

1.2.4 Метод прогонки

Пусть A — трехдиагональная матрица, которую мы представим в виде:

$$A = \begin{pmatrix} c_1 & -b_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -a_2 & c_2 & -b_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -a_3 & c_3 & -b_3 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -a_N & c_N \end{pmatrix}$$

Знак $-$ перед b_i, c_i поставлен для удобства. Для решения задачи $A\mathbf{t} = \mathbf{s}$ в этом случае применяется *метод прогонки.*

Положим $a_1 = b_N = 0$, тогда трехдиагональная система может быть записана в виде

$$-t_{k-1}a_k + t_k c_k - t_{k+1}b_k = s_k \ , \quad k = 1, 2, \ldots, N \ .$$

Рассмотрим эту систему подробнее. Выразим из первого уравнения t_1 через t_2:

$$t_1 c_1 - t_2 b_1 = s_1 \quad \Rightarrow \quad t_1 = \frac{b_1}{c_1}t_2 + \frac{s_1}{c_1} \ .$$

Теперь из второго уравнения выразим t_2 через t_3: $-t_1 a_2 + t_2 c_2 - t_3 b_2 = s_2$, или

$$-\left(\frac{s_1}{c_1} + \frac{b_1}{c_1}t_2\right)a_2 + t_2 c_2 - t_3 b_2 = s_2 \quad \Rightarrow \quad t_2 = \frac{b_2 t_3}{c_2 - \frac{b_1}{c_1}a_2} + \frac{s_2 + \frac{a_2}{c_1}s_1}{c_2 - \frac{b_1}{c_1}a_2} \ .$$

Аналогично,

$$t_k = \alpha_k t_{k+1} + \beta_k \ ,$$

где

$$\alpha_k = \frac{b_k}{c_k - \alpha_{k-1}a_k} \quad , \quad \beta_k = \frac{s_k + \beta_{k-1}a_k}{c_k - \alpha_{k-1}a_k} \ .$$

Убедимся в справедливости этого представления по индукции. Действительно, $\alpha_1 = \frac{b_1}{c_1}$, $\beta_1 = \frac{s_1}{c_1}$, таким образом база индукции верна. Теперь осуществим собственно индукционный переход. Пусть $t_k = \alpha_k t_{k+1} + \beta_k$, тогда

$$-a_{k+1}t_k + c_{k+1}t_{k+1} - b_{k+1}t_{k+2} = s_{k+1} \ ,$$

$$-a_{k+1}(\alpha_k t_{k+1} + \beta k) + c_{k+1}t_{k+1} - b_{k+1}t_{k+2} = s_{k+1} \ ,$$

откуда

$$t_{k+1} = \frac{b_{k+1}t_{k+2}}{c_{k+1} - \alpha_k a_{k+1}} + \frac{s_{k+1} + \beta_k a_{k+1}}{c_{k+1} - \alpha_k a_{k+1}} = \alpha_{k+1}t_{k+2} + \beta_{k+1} \ ,$$

то есть индукционный переход также имеет место.

Рассмотрим теперь каким образом применяется метод прогонки. На первом этапе (прямой ход прогонки) мы определяем коэффициенты α_k, β_k через известные нам элементы матрицы A (b_k, c_k, a_k), заданные значения s_k и предыдущие α_{k-1}, β_{k-1}:

$\alpha_1 = \frac{b_1}{c_1}$, $\beta_1 = \frac{s_1}{c_1}$, начало прямого хода,

$\alpha_k = \frac{b_k}{c_k - \alpha_{k-1}a_k}$, $\beta_k = \frac{s_k + \beta_{k-1}a_k}{c_k - \alpha_{k-1}a_k}$, прямой ход.

После того как определены коэффициенты α_k и β_k начинается обратный ход прогонки — собственно определение компонент t_k. Имеем

$$t_N = \alpha_N t_{N+1} + \beta_N \ ,$$

при этом $\alpha_N = 0$, т.к. $b_N = 0$, а $\alpha_N = \frac{b_N}{c_N - \alpha_{N-1}a_N}$. Таким образом,

$t_N = \beta_N$ (начало обратного хода) ,

$t_k = \alpha_k t_{k+1} + \beta_k$ (обратный ход) .

Утверждение (Достаточное условие разрешимости прогонки): *Пусть* $|c_k| > |b_k| + |a_k|$, $k = 1, \ldots, N$, *тогда* $\det A \neq 0$.

Доказательство. Необходимо убедиться, что знаменатель в формулах прямого хода не обращается в нуль. Для этого достаточно убедиться в том, что $|\alpha_k| < 1$. Ведь если это так, то

$$|c_k - \alpha_{k-1} a_k| \geq |c_k| - |\alpha_{k-1}||a_k| > |c_k| - |a_k| > |b_k| \geq 0$$

и не происходит деления на нуль. Имеем :

$$|\alpha_1| = |\frac{b_1}{c_1}| < 1 \ , \quad |\alpha_k| = \frac{|b_k|}{|c_k - \alpha_{k-1} a_k|} < \frac{|b_k|}{|b_k|} = 1 \ .$$

1.2.5 Метод итераций для решения линейных систем

Система линейных уравнений $A\mathbf{x} = \mathbf{b}$:

$$\sum_{j=1}^{N} a_{ij} x_j = b_i \ , \quad i = 1, 2, \ldots, N \ , \tag{1}$$

может быть решена не только прямыми методами, но также и итерационными. Разумеется, мы предполагаем, что система имеет единственное решение, т.е., что $\det A \neq 0$.

Представим матрицу A в виде $A = B + D$, где $D = \mathrm{diag}\{a_{11}, \ldots, a_{NN}\}$. Предположим, что $\det D \neq 0$, что равносильно тому, что $a_{ii} \neq 0$, $i = 1, \ldots, N$ (если исходно это не так, то перестановкой строк и столбцов этого всегда можно добиться при $\det A \neq 0$). Тогда (1) переписывается в виде $D\mathbf{x} = \mathbf{b} - B\mathbf{x}$, или

$$\mathbf{x} = D^{-1}\mathbf{b} - D^{-1}B\mathbf{x} \ .$$

Предложим следующую итерационную процедуру

$$\mathbf{x}^{s+1} = D^{-1}\mathbf{b} - D^{-1}B\mathbf{x}^s \ ,$$

$\mathbf{x}^0$ — произвольный начальный вектор. В развернутой форме

$$x_i^{s+1} = a_{ii}^{-1} b_i - a_{ii}^{-1} \sum_{j=1, j\neq i}^{n} a_{ij} x_j^s \ , \quad i = 1, 2, \ldots N \ .$$

Обозначим $D^{-1}\mathbf{b} = \mathbf{u}$, $D^{-1}B = T$, тогда итерационный процесс принимает вид

$$\mathbf{x}^{s+1} = \mathbf{u} - T\mathbf{x}^s \ . \tag{2}$$

Теорема1. *Процесс (2) сходится, для любого начального вектора, если* $||D^{-1}(A - D)|| = ||T|| < 1$.

Доказательство. Для доказательства достаточно заметить, что отображение $\mathbf{x} \to \mathbf{u} - T\mathbf{x}$ является сжатием.

Таким образом, последовательность $\mathbf{x}^s$ имеет предел. Пусть $\mathbf{x}^* = \lim_{s\to\infty} \mathbf{x}^s$, тогда $\mathbf{x}^* = \mathbf{u} - T\mathbf{x}^*$, или, возвращаясь к исходной формулировке, $A\mathbf{x}^* = \mathbf{b}$. Итак, для сходимости изложенного метода, называемого *методом простых итераций*, необходимо чтобы

$$||D^{-1}(A - D)|| < 1 \ .$$

1.2.6 Метод Зейделя

Модифицируем метод простых итераций, координатную форму которого, в частности, можно записать в виде

$$x_i^{s+1} = a_{ii}^{-1}[b_i - \underbrace{\sum_{j<i} a_{ij}x_j^s}_{*} - \sum_{j>i} a_{ij}x_j^s] \ , \quad i = 1,2,\ldots,N \ .$$

Заметим, что если последовательно вычислять компоненты $(s+1)$-го приближения $\mathbf{x}^{s+1}$, начиная с первой x_1^{s+1}, то к моменту вычисления конкретной i-ой компоненты x_i^{s+1}, координаты $x_1^{s+1}, \ldots, x_{i-1}^{s+1}$ уже определены и их можно было бы использовать для определения более точного последующего приближения $\mathbf{x}^{s+1}$. Модифицируем соответствующим образом метод простых итераций, заменив в сумме $*$ компоненты x_j^s на x_j^{s+1}. Таким образом, мы получаем новую итерационную процедуру

$$x_i^{s+1} = a_{ii}^{-1}b_i - a_{ii}^{-1}\sum_{j<i} a_{ij}x_j^{s+1} - a_{ii}^{-1}\sum_{j>i} a_{ij}x_j^s \ , \quad i = 1,2,\ldots,N \ .$$

Такой итерационный процесс называется *методом Зейделя.* Представим его в матричной форме. Пусть L — нижнетреугольная матрица с элементами

L :

$$l_{ij} = \begin{cases} a_{ij} & , \ j < i \\ 0 & , \ j \geq i \end{cases} \ ,$$

а U — верхнетреугольная матрица с элементами

$$u_{ij} = \begin{cases} a_{ij} & , \ j > i \\ 0 & , \ j \leq i \end{cases} \ .$$

Как и раньше введем матрицу $D = diag\{a_{11} \ldots a_{NN}\}$, тогда $A = D + L + U$. В матричном виде метод Зейделя имеет вид:

$$\mathbf{x}^{s+1} = D^{-1}\mathbf{b} - D^{-1}L\mathbf{x}^{s+1} - D^{-1}U\mathbf{x}^s \ .$$

Сходимость метода Зейделя

Итак, итерации по методу Зейделя должны быть организованы таким образом, чтобы

$$D\mathbf{x}^{s+1} = \mathbf{b} - L\mathbf{x}^{s+1} - U\mathbf{x}^s \ ,$$

или

$$\mathbf{x}^{s+1} = (D+L)^{-1}\mathbf{b} - (D+L)^{-1}U\mathbf{x}^s \ .$$

Отображение $\mathbf{x} \mapsto (D+L)^{-1}\mathbf{b} - (D+L)^{-1}U\mathbf{x}$ является сжатием, если $||(D+L)^{-1}U|| < 1$, то есть справедлива

Теорема. *Метод Зейделя сходится, если* $||(D+L)^{-1}U|| < 1$.

Условия этой теоремы довольно трудно проверяемы, так как матрица $(D+L)^{-1}U$ должна еще и вычисляться. Существует достаточно простой признак сходимости метода Зейделя, который связан с понятием

положительной определенности матрицы относительно скалярного произведения. Напомним, что оператор A, действующий в евклидовом пространстве E_n называется положительно определенным, если

$$\langle A\mathbf{x}, \mathbf{x} \rangle \geq \gamma \langle \mathbf{x}, \mathbf{x} \rangle \ , \ \gamma > 0 \ .$$

Если оператор положительно определен, то у него существует и обратный и он также положительно определен. Также важно отметить, что если оператор A положительно определен и симметричен в R^N, то форма

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \langle A\mathbf{x}, \mathbf{y} \rangle$$

удовлетворяет всем свойствам скалярного произведения. В дальнейшем, факт положительной определенности оператора A будем обозначать: $A > 0$. Заметим, что в комплексном евклидовом пространстве положительная определенность оператора A автоматически влечет за собой эрмитовость: $A = A^*$.

Теорема (достаточный признак сходимости метода Зейделя). *Метод Зейделя сходится в вещественном евклидовом пространстве, если A симметричная положительно определенная матрица.*

Для доказательства этой теоремы нам потребуется следующая

Лемма. *Пусть последовательность векторов $\mathbf{z}^k$ в $\mathbf{R}^N$ определена рекуррентным соотношением*

$$B(\mathbf{z}^{k+1} - \mathbf{z}^k) + A\mathbf{z}^k = 0 \ , \tag{3}$$

где $B - \frac{1}{2}A > 0$, $A > 0$ и симметрична, тогда $\mathbf{z}^k \to 0$.

Доказательство. Представим $\mathbf{z}^k$ в виде

$$\mathbf{z}^k = \frac{1}{2}(\mathbf{z}^{k+1} + \mathbf{z}^k) - \frac{1}{2}(\mathbf{z}^{k+1} - \mathbf{z}^k) \ ,$$

и подставим это представление в (3), тогда

$$B(\mathbf{z}^{k+1} - \mathbf{z}^k) + \frac{1}{2}A(\mathbf{z}^{k+1} + \mathbf{z}^k) - \frac{1}{2}A(\mathbf{z}^{k+1} - \mathbf{z}^k) = 0 \ ,$$

или

$$(B - \frac{1}{2}A)(\mathbf{z}^{k+1} - \mathbf{z}^k) + \frac{1}{2}A(\mathbf{z}^{k+1} + \mathbf{z}^k) = 0 \ .$$

Умножим это равенство скалярно на $\mathbf{z}^{k+1} - \mathbf{z}^k$, тогда

$$0 = |\mathbf{z}^{k+1} - \mathbf{z}^k|_{B-A/2} + \frac{1}{2}\langle A(\mathbf{z}^{k+1} + \mathbf{z}^k), \mathbf{z}^{k+1} - \mathbf{z}^k \rangle$$

$$= |\mathbf{z}^{k+1} - \mathbf{z}^k|_{B-A/2} + \frac{1}{2}\{|\mathbf{z}^{k+1}|_A - |\mathbf{z}^k|_A\} = 0 \ ,$$

где $|\cdot|_A = \langle A\cdot, \cdot \rangle$, $|\cdot|_{B-A/2} = \langle \{B - A/2\}\cdot, \cdot \rangle$ — нормы, определяемые операторами A и $B - A/2$, соответственно. Из последнего равенства в силу положительной определенности оператора $(B - A/2)$ следует, что $|\mathbf{z}^{k+1}|_A - |\mathbf{z}^k|_A \leq 0$, т.е. последовательность $|\mathbf{z}^k|_A$ невозрастающая: $|\mathbf{z}^{k+1}|_A \leq |\mathbf{z}^k|_A$. При этом, последовательность чисел $|\mathbf{z}^k|_A$ ограничена снизу поскольку $|\mathbf{z}^k|_A \geq 0$. Таким образом, существует конечный предел $\lim\limits_{k\to\infty} |\mathbf{z}^k|_A = a$. Но тогда из того же равенства следует, что норма $|\mathbf{z}^{k+1} - \mathbf{z}^k|_{(B-\frac{1}{2}A)}$ стремится к нулю, а значит и $\mathbf{z}^{k+1} - \mathbf{z}^k \to 0$, $k \to \infty$. Вернемся теперь к определению последовательности $\mathbf{z}^k$:

$$A\mathbf{z}^k = -B(\mathbf{z}^{k+1} - \mathbf{z}^k) \ ,$$

откуда $\mathbf{z}^k = -A^{-1}B(\mathbf{z}^{k+1} - \mathbf{z}^k)$ и, следовательно,

$$||\mathbf{z}^k|| \leq ||A^{-1}B|| \times ||\mathbf{z}^{k+1} - \mathbf{z}^k|| \to 0 \ ,$$

и, таким образом, $\mathbf{z}^k \to 0$, при $k \to \infty$.

Приступим теперь собственно к доказательству достаточного признака сходимости метода Зейделя. Как нетрудно видеть, метод Зейделя $(D+L)\mathbf{x}^{s+1} + U\mathbf{x}^s = \mathbf{b}$ может быть представлен в виде

$$(D+L)(\mathbf{x}^{s+1} - \mathbf{x}^s) + A\mathbf{x}^s = \mathbf{b} .$$

Пусть $\mathbf{u}$ — точное решение уравнения $A\mathbf{u} = \mathbf{b}$, оно существует, так как A — положительно определенный оператор и, следовательно, обратим. Положим также $\mathbf{z}^s = \mathbf{x}^s - \mathbf{u}$, тогда

$$(D+L)(\mathbf{z}^{s+1} - \mathbf{z}^s) + A\mathbf{z}^s = 0 .$$

Убедимся в том, что $(D + L - \frac{1}{2}A)$ положительно определенная матрица, если A симметрична и положительно определена. Действительно

$$D + L - \frac{1}{2}A = D + L - \frac{1}{2}(D + L + U) = \frac{1}{2}(D + L - U) .$$

Рассмотрим соответствующую квадратичную форму

$$\langle (D + L - U)\mathbf{x}, \mathbf{x}\rangle = \langle D\mathbf{x}, \mathbf{x}\rangle + \langle L\mathbf{x}, \mathbf{x}\rangle - \langle U\mathbf{x}, \mathbf{x}\rangle .$$

Заметим, что поскольку A — симметричная матрица, следовательно $L^T = U$ и

$$\langle L\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{x}, L^T\mathbf{x}\rangle = \langle \mathbf{x}, U\mathbf{x}\rangle = \langle U\mathbf{x}, \mathbf{x}\rangle ,$$

поэтому

$$\langle (D + L - U)\mathbf{x}, \mathbf{x}\rangle = \langle D\mathbf{x}, \mathbf{x}\rangle = \sum_i a_{ii} x_i^2 > 0 ,$$

так как у положительно определенной матрицы все диагональные элементы больше нуля (почему?): $a_{ii} > 0$. Таким образом, мы находимся в условиях Леммы, следовательно, последовательность $\mathbf{z}^s$ стремится к нулю, откуда следует, что последовательность $\mathbf{x}^s = \mathbf{u} + \mathbf{z}^s$ стремится к истинному решению $\mathbf{u}$.

Глава 2

Алгебраические спектральные задачи

2.1 Некоторые сведения из матричной теории

Пусть A — линейный оператор, действующий в вещественном $\mathbf{R}^N$ или в комплексном $\mathbf{C}^N$ евклидовом пространстве: $A:\mathbf{R}^N(\mathbf{C}^N)\to\mathbf{R}^N(\mathbf{C}^N)$.

Число λ и вектор $\mathbf{x}$ называются, соответственно, *собственным числом (значением)* и *собственным вектором* оператора A, отвечающим собственному числу λ , если $A\mathbf{x}=\lambda\mathbf{x}$.

В частности, справедливы следующие теоремы.

Теорема 1.*Всякий линейный оператор в* $\mathbf{C}^N$ *имеет по крайней мере одно собственное значение.*

Теорема 2. *Собственные векторы, отвечающие различным собственным значениям, линейно независимы.*

Теорема 3. *Для любого набора из* N *линейно независимых векторов* $\mathbf{e}^1,\mathbf{e}^2,\dots,\mathbf{e}^N$ *(базиса) существует единственный дуальный базис* $\tilde{\mathbf{e}}^1,\tilde{\mathbf{e}}^2,\dots,\tilde{\mathbf{e}}^N$ *, такой что* $\langle\mathbf{e}^i,\tilde{\mathbf{e}}^j\rangle=\delta_{ij}$.

Заметим, что всякий ортонормированный базис самодуален. Пусть A имеет N различных собственных векторов $\mathbf{x}^i$, тогда они образуют базис, и, следовательно, существует дуальный базис $\tilde{\mathbf{x}}^i$. В этом случае, как нетрудно убедиться, сопряженный оператор A^* (в случае вещественного евклидова пространства просто транспонированная матрица A^T) имеет в качестве собственных значений числа λ_i , а в качестве собственных векторов — векторы дуального базиса:

$$A\mathbf{x}^i=\lambda_i\mathbf{x}^i\ ,\quad A^*\tilde{\mathbf{x}}^i=\lambda_i\tilde{\mathbf{x}}^i\ .$$

Действительно: $\langle\mathbf{x}^i,\lambda_i\tilde{\mathbf{x}}^i\rangle=\langle\lambda\mathbf{x}^i,\tilde{\mathbf{x}}^i\rangle=\langle A\mathbf{x}^i,\tilde{\mathbf{x}}^i\rangle=\langle\mathbf{x}^i,A^*\tilde{\mathbf{x}}^i\rangle$ и, аналогично, $\langle\mathbf{x}^j,A^*\tilde{\mathbf{x}}^i\rangle=0$, $i\neq j$, то есть $\langle\mathbf{x}^j,A^*\tilde{\mathbf{x}}^i\rangle=\delta_{ij}\langle\mathbf{x}^j,\lambda\tilde{\mathbf{x}}^j\rangle$. Кроме того, нетрудно показать справедливость следующего спектрального разложения оператора A :

$$A\cdot=\sum_{i=1}^{N}\lambda_i\langle\cdot,\tilde{\mathbf{x}}^i\rangle\mathbf{x}^i=\sum_{i=1}^{N}\lambda_iP_i\cdot\ ,$$

где операторы $P_i\cdot=\langle\cdot,\tilde{\mathbf{x}}^i\rangle\mathbf{x}^i$ суть *собственные проекторы* оператора A . В самом деле, произвольный вектор $\mathbf{f}$ можно разложить по собственным векторам оператора A : $\mathbf{f}=\sum\langle\mathbf{f},\tilde{\mathbf{x}}^i\rangle\mathbf{x}^i$. Тогда

$$A\mathbf{f}=\sum\langle\mathbf{f},\tilde{\mathbf{x}}^i\rangle A\mathbf{x}^i=\sum\langle\mathbf{f},\tilde{\mathbf{x}}^i\rangle\lambda_i\mathbf{x}^i=\sum\lambda_iP_i\mathbf{f}\ .$$

Пусть $A = A^*$ — эрмитова матрица (в $\mathbf{R}^N$ — симметричная). В этой ситуации собственные значения вещественны, алгебраическая и геометрическая кратности любого собственного значения совпадают, собственные векторы $\mathbf{x}^i$, отвечающие различным собственным числам ортогональны, и существует ортонормированный базис из собственных векторов. В случае однократно вырожденного собственного значения λ, отвечающий ему собственный проектор P одномерен и имеет вид $P = \langle\cdot, \mathbf{x}\rangle\mathbf{x}$ (всюду считаем, что собственный вектор $\mathbf{x}$ нормирован на единицу). Если подпространство решений $A\mathbf{x} = \lambda\mathbf{x}$ более чем одномерно, то в нем выбирается произвольный ортобазис $\mathbf{x}^i$ и собственный проектор, отвечающий собственному числу λ , представляет собой сумму соответствующих одномерных проекторов $P = \sum\langle\cdot, \mathbf{x}^i\rangle\mathbf{x}^i$.

Отметим (легко проверяемое) важное свойство ортогональных проекторов:

$$P_i P_k = \delta_{ik} P_i \ .$$

Степень оператора имеет следующую запись через ортогональные проекторы

$$A^m = \sum_k \lambda_k^m(A) P_k \ .$$

Многочлены от оператора определяются как сумма соответствующих степеней. Поскольку многочленами можно приблизить любую функцию, то функцию от оператора естественно определить как

$$f(A) = \sum_k f(\lambda_k) P_k \ .$$

Собственные функции у оператора и у функции от оператора совпадают, тогда как собственные значения функции от оператора есть числа $f(\lambda_k)$.

2.2 Собственные числа эрмитовых матриц

2.2.1 Интерполяционный метод

Поскольку собственные числа λ_i матрицы A являются корнями характеристического полинома $F_A(\lambda) = \det(A - \lambda I)$, то можно вычислить $F_A(\lambda)$ в $(n+1)$-ом наугад выбранном значении λ (их естественно выбирать в промежутке $(-||A||, ||A||)$, если границы спектра известны; оценить их можно по максимальному по модулю элементу матрицы) и построить по ним интерполяционный полином степени n, который совпадает собственно с характеристическим, после чего определяются его корни. Этот метод применим и для неэрмитовых матриц (при соответствующем выборе метода определения корней).

2.2.2 Нахождение максимального по модулю собственного значения

Для удобства будем считать, что собственные числа пронумерованы в порядке убывания их модуля.

а) Метод итераций

Пусть $\mathbf{g} = \mathbf{g}^0$ — произвольный начальный вектор. Определим последовательность

$$\mathbf{g}^n = A\frac{\mathbf{g}^{n-1}}{||\mathbf{g}^{(n-1)}||} = \frac{A^n\mathbf{g}}{||A^{n-1}\mathbf{g}||} \ ,$$

тогда

$$\lim ||\mathbf{g}^{(n)}|| = |\lambda_{max}| \ .$$

Доказательство: Действительно, $\mathbf{g} = \sum P_k \mathbf{g}$, $||\mathbf{g}^n|| = \frac{||A^n \mathbf{g}||}{||A^{n-1}\mathbf{g}||}$, и

$$A^n \mathbf{g} = \sum \lambda_k^n P_k \mathbf{g} = \lambda_{max}^n (P_{max}\mathbf{g} + \sum_{k \neq 1} \{\lambda_k/\lambda_{max}\}^n P_k \mathbf{g}) \ .$$

Пусть λ' – собственное число, следующее за максимальным по модулю. Тогда

$$||A^n \mathbf{g}||^2 = \langle A^n \mathbf{g}, A^n \mathbf{g} \rangle = \lambda_{max}^{2n} (\langle P_{max}\mathbf{g}, \mathbf{g} \rangle + O(|\lambda'/\lambda_{max}|^{2n})) \ ,$$

и

$$||\mathbf{g}^n|| = \frac{||A^n g||}{||A^{n-1}\mathbf{g}||} = |\lambda_{max}| \sqrt{\frac{\langle P_{max}\mathbf{g}, \mathbf{g} \rangle + O(|\lambda'/\lambda_{max}|^{2n})}{\langle P_{max}\mathbf{g}, \mathbf{g} \rangle + O(|\lambda'/\lambda_{max}|^{2n-2})}}$$
$$= |\lambda_{max}| \{1 + O(|\lambda'/\lambda_{max}|^{2n-2})\} \ .$$

Таким образом, если стартовый вектор $\mathbf{g}$ имел ненулевую проекцию на собственное подпространство, отвечающее максимальному по модулю собственному значению (то есть $P_{max}\mathbf{g} \neq 0$), то приведенная итерационная процедура приводит к нахождению λ_{max} . Однако, хотя формально, предыдущее рассмотрение верно лишь в случае ненулевой проекции, в действительности, из-за ошибок округления при вычислениях эта проекция наверняка появится на некотором шаге и дальнейшее применение метода итераций приведет к желаемому результату. Попутно заметим, что если подпространство отвечающее λ_{max} одномерно, то метод итераций одновременно приводит к нахождению собственного вектора $\mathbf{x}^{max}$, отвечающего λ_{max} . Этим вектором с точностью до нормировки является

$$\mathbf{x}^{max} = \lim_{n \to \infty} \mathbf{g}^n \ .$$

Замечание. Для нахождения λ_{max} можно применять метод итераций и в более простой постановке. Пусть l-ая компонента в максимального собственного вектора в стандартном евклидовом базисе не равна нулю (хотя бы одна такая существует), тогда

$$\lambda_{max} = \lim_{n \to \infty} \frac{(A^n \mathbf{g})_l}{(A^{n-1}\mathbf{g})_l} \ .$$

б) Метод следов

Известно, что след матрицы (сумма диагональных элементов) равен сумме её собственных значений с учетом кратности: $\sum \lambda_i = TrA$, таким образом, $\sum \lambda_i^m = TrA^m$ и, следовательно,

$$TrA^m = \lambda_{max}^m [1 + (\lambda'/\lambda_{max})^m + \ldots] \ ,$$

где λ' – следующее по модулю за максимальным собственное значение. Таким образом, λ_{max} можно искать как следующий предел

$$|\lambda_{max}| = \lim_{m \to \infty} \sqrt[m]{TrA^m} \ ,$$

или, например, в виде

$$\lambda_{max} = \lim_{m \to \infty} \frac{TrA^{m+1}}{TrA^m} \ .$$

Процедуру возведения матрицы в степень можно оптимизировать:

$$\underbrace{\underbrace{A \times A}_{A^2} \times \underbrace{A \times A}_{A^2}}_{A^4} ,$$

и так далее, в частности: $A^{16} = (A^8)^2 = A^{2^{2^{2^2}}}$.

в) Метод скалярных произведений

Этот метод является обобщением метода итераций. Пусть $\mathbf{x}$ и $\mathbf{y}$ произвольные начальные векторы. Определим итерации

$$\mathbf{y}^m = A\mathbf{y}^{m-1} = A^m\mathbf{y} = \sum_s \lambda^m P_s \mathbf{y} \ ,$$

$$\mathbf{x}^m = A\mathbf{x}^{m-1} = \sum_s \lambda^m P_s \mathbf{x} \ .$$

Аналогично методу итераций убеждаемся, что

$$\frac{\langle \mathbf{y}^m, \mathbf{x}^m \rangle}{\langle \mathbf{y}^m, \mathbf{x}^{m-1} \rangle} \to \lambda_{max} \ .$$

2.2.3 Обратные итерации

Поиск минимального по модулю собственного значения

Пусть $\mathbf{y}$ — некоторый стартовый вектор. Определим обратные итерации как $\mathbf{y}^{(n)} = A\mathbf{y}^{(n+1)}$ (или $\mathbf{y}^{(n+1)} = A^{-1}\mathbf{y}^{(n)}$) , то есть это прямая задача для нахождения максимального собственного значения μ_{max} матрицы $B = A^{-1}$, обратной к исходной матрице. Очевидно, что минимальное по модулю собственное значение матрицы A равно максимальному по модулю собственному числу обратной матрицы.

$$\begin{matrix} B & = & A^{-1} \\ \mu_i & & \lambda_i \end{matrix} \quad , \quad (\mu_i = \frac{1}{\lambda_i}) \ .$$

Метод обратных итераций со сдвигом

Пусть A — невырожденная эрмитова матрица и λ_* — некоторое пробное число. Рассмотрим матрицу $(A-\lambda_* I)$, ее собственными значениями являются числа $(\lambda_i - \lambda_*)$, где λ_i — собственные значения исходной матрицы A . У обратной матрицы $(A-\lambda_* I)^{-1}$ собственные значения — это величины $\frac{1}{\lambda_i - \lambda_*}$. Процедура метода обратных итераций со сдвигом

$$\mathbf{y}^{(n)} = (A - \lambda_* I)\mathbf{y}^{(n+1)} \ ,$$

приводит к нахождению $\max_i |\frac{1}{\lambda_i - \lambda_*}|$. Иными словами, мы находим то собственное значение λ_j , которое является ближайшим к пробному числу λ_* . Варьируя пробное λ_* и вновь применяя метод обратных итераций со сдвигом можно, найти все собственные значения матрицы A .

2.3 Неэрмитовы матрицы

2.3.1 Дополнительные сведения

В случае если алгебраическая и геометрическая кратности собственных чисел оператора A совпадают, то унитарным преобразованием (то есть преобразованием, сохраняющим скалярное произведение: $\langle U\mathbf{x}, U\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$) (в $\mathbf{R}^N$ — ортогональным преобразованием) оператор приводится к диагональному виду, и на диагонали стоят собственные числа A с учетом кратности. Однако нередка ситуация, когда алгебраическая кратность собственного значения превышает геометрическую (обратное, кстати, невозможно).

В $\mathbf{C}^N$ при определенном выборе базиса (называемым жордановым или каноническим базисом оператора A) матрица оператора становится блочно-диагональной. В каждом из блоков (жордановых клеток) матрица оператора является верхнетреугольной и имеет вид

$$\begin{pmatrix} \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & \dots & 0 & 0 \\ 0 & 0 & \lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{pmatrix} . \tag{1}$$

Размеры жордановых клеток, их количество, также как и числа λ (корни характеристического уравнения) являются инвариантами оператора A (то есть не зависят от выбора жорданова базиса).

В $\mathbf{R}^N$ жорданов базис приводит к клеткам вида (1) если λ — вещественный корень характеристического уравнения матрицы оператора A в каком либо базисе. Поскольку коэффициенты характеристического полинома матрицы оператора в $\mathbf{R}^N$ вещественны, то вместе с каждым комплексным корнем $\lambda = \mu + i\nu$ он обладает и комплексно сопряженным $\lambda = \mu - i\nu$. Жорданова клетка в этом случае имеет вид

$$\begin{pmatrix} \mu & \nu & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ -\nu & \mu & 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \mu & \nu & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & -\nu & \mu & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & \nu & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & -\nu & \mu & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \mu & \nu \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & -\nu & \mu \end{pmatrix} .$$

2.3.2 Метод итераций для максимального по модулю собственного числа кратности 2 в случае жордановой аномалии

Остановимся подробно на случае, когда максимальному по модулю собственному значению λ оператора A соответствует жорданова клетка размера 2×2 . В каноническом базисе $\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^N$ матрица оператора A имеет вид

$$\begin{pmatrix} \lambda & 1 & | & 0 & \dots & 0 \\ 0 & \lambda & | & 0 & \dots & 0 \\ - & - & - & - & - & - \\ 0 & 0 & | & & & \\ \vdots & \vdots & | & & \mathrm{B} & \\ 0 & 0 & | & & & \end{pmatrix} .$$

Здесь B — матрица, отвечающая оставшимся собственным значениям, конкретный вид которой нас не интересует. Обозначим дуальный базис через $\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^N$. Тогда

$$\begin{array}{ll} A\mathbf{u}^1 = \lambda\mathbf{u}^1 & A^*\mathbf{v}^1 = \lambda\mathbf{v}^1 \\ A\mathbf{u}^2 = \lambda\mathbf{u}^2 + \mathbf{u}^1 & A^*\mathbf{v}^2 = \lambda\mathbf{v}^2 + \mathbf{v}^1 \end{array} .$$

Вектор $\mathbf{u}^1$ является собственным для оператора A, соответствующим собственному значению λ. Вектор $\mathbf{u}^2$ называется присоединенным. Для сопряженного оператора A^* собственным и присоединенным векторами, соответствующими собственному значению λ (в $\mathbf{R}^N$ просто λ) являются векторы дуального базиса $\mathbf{v}^1$ и $\mathbf{v}^2$, соответственно. Заметим, что $\mathbf{u}^1 \quad \mathbf{v}^2$, и $\mathbf{u}^2 \quad \mathbf{v}^1$, то есть собственный вектор для оператора является присоединенным для сопряженного и наоборот.

Непригодность обычного метода итераций

Будем считать, что собственные значения пронумерованы в порядке убывания модуля и что $\lambda_1 = \lambda$. Пусть $\mathbf{x}$ произвольный вектор. Разложим его по векторам жорданова базиса и дуального к нему

$$\mathbf{x} = \sum_{i=1}^{N} \langle \mathbf{x}, \mathbf{v}^i \rangle \mathbf{u}^i \ , \quad \mathbf{x} = \sum_{i=1}^{N} \langle \mathbf{u}^i, \mathbf{x} \rangle \mathbf{v}^i \ .$$

Подействуем на $\mathbf{x}$ оператором A и сопряженным:

$$A\mathbf{x} = \sum_{i=1}^{N} \langle \mathbf{x}, \mathbf{v}^i \rangle A\mathbf{u}^i \ , \quad A^*\mathbf{x} = \sum_{i=1}^{N} \langle \mathbf{u}^i, \mathbf{x} \rangle A^*\mathbf{v}^i \ ,$$

или

$$A\mathbf{x} = (\lambda < \mathbf{x}, \mathbf{v}^1 > + < \mathbf{x}, \mathbf{v}^2 >)\mathbf{u}^1 + \lambda < \mathbf{x}, \mathbf{v}^2 > \mathbf{u}^2 + \sum_{i=3}^{N} < \mathbf{x}, \mathbf{v}^i > A\mathbf{u}^i \ ,$$

$$A^*\mathbf{x} = (\lambda < \mathbf{u}^1, \mathbf{x} > + < \mathbf{u}^2, \mathbf{x} >)\mathbf{v}^1 + \lambda < \mathbf{u}^2, \mathbf{x} > \mathbf{v}^2 + \sum_{i=3}^{N} < \mathbf{u}^i, \mathbf{x} > A^*\mathbf{v}^i \ .$$

Аналогично, поскольку

$$\mathrm{A}^n = \begin{pmatrix} \lambda^n & n\lambda^{n-1} & | & 0 & \dots & 0 \\ 0 & \lambda^n & | & 0 & \dots & 0 \\ - & - & - & - & - & - \\ 0 & 0 & | & & & \\ \vdots & \vdots & | & & \mathrm{B}^n & \\ 0 & 0 & | & & & \end{pmatrix} ,$$

то

$$A^n\mathbf{x} = (\lambda^n < \mathbf{x}, \mathbf{v}^1 > + n\lambda^{n-1} < \mathbf{x}, \mathbf{v}^2 >)\mathbf{u}^1 + \lambda^n < \mathbf{x}, \mathbf{v}^2 > \mathbf{u}^2 + \dots \ . \qquad (2)$$

Квадратичная форма n-ой степени оператора A с использованием (2) может быть записана как

$$\langle A^n\mathbf{x}, \mathbf{x} \rangle = \langle A^n\mathbf{x}, \sum_{1}^{N} \langle \mathbf{u}^i, \mathbf{x} \rangle \mathbf{v}^i \rangle =$$

$$\lambda^n \left\{ \left(a + b\frac{n}{\lambda} \right) + O\left(|\lambda'/\lambda|^n \right) \right\} ,$$

где $a = < \mathbf{x}, \mathbf{v}^1 >< \mathbf{u}^1, \mathbf{x} > + < \mathbf{x}, \mathbf{v}^2 >< \mathbf{u}^2, \mathbf{x} >$ ($= 2 < \mathbf{x}, \mathbf{v}^1 >< \mathbf{u}^1, \mathbf{x} >$ в $\mathbf{R}^n$), $b = < \mathbf{x}, \mathbf{v}^2 >< \mathbf{u}^1, \mathbf{x} >$ и λ' — следующее по модулю за λ собственное значение.

В нашей ситуации λ вещественно. По аналогии с методом скалярных произведений, применяемым для эрмитовых матриц, рассмотрим отношение Релея

$$\rho_n = \frac{\langle A^{n+1}\mathbf{x}, A^{*n}\mathbf{x} \rangle}{\langle A^n\mathbf{x}, A^{*n}\mathbf{x} \rangle} = \frac{\langle A^{2n+1}\mathbf{x}, \mathbf{x} \rangle}{\langle A^{2n}\mathbf{x}, \mathbf{x} \rangle} = \lambda \left\{ \frac{a + \frac{(2n+1)b}{\lambda}}{a + \frac{2nb}{\lambda}} + O\left(|\lambda'/\lambda|^{2n} \right) \right\}$$

$$\lambda\left\{1+\frac{a+\frac{b}{\lambda}}{a+\frac{2nb}{\lambda}}+O\left(|\lambda'/\lambda|^{2n}\right)\right\}=\lambda\left\{1+O\left(\frac{1}{n}\right)\right\} .$$

Итак, $\rho_n=\lambda\{1+O(1/n)\}$, то есть сходимость при $n\to\infty$ настолько неудовлетворительная, что теряет практический смысл. Таким образом, обычными итерационными методами собственное число в случае жордановой аномалии сосчитать не представляется возможным. Необходим какой-то другой подход.

Модифицированный метод итераций

Составим квадратное уравнение, для которого λ является корнем кратности 2

$$(t-\lambda)^2=t^2+pt+q=0 \quad p=-2\lambda\ ,\ q=\lambda^2\ .$$

Коэффициенты p и q заранее неизвестны, поскольку неизвестно само λ . Попытаемся их определить. Обозначим $\mathbf{x}^n=A^n\mathbf{x}$ и рассмотрим выражение

$$\mathbf{x}^{n+1}+p\mathbf{x}^n+q\mathbf{x}^{n-1}=<\mathbf{x},\mathbf{v}^1>\underbrace{\{\lambda_1^{n+1}+p\lambda_1^n+q\lambda_1^{n-1}\}}_{0}\mathbf{u}^1+$$

$$+<\mathbf{x},\mathbf{v}^2>\{(n+1)\lambda_1^n+pn\lambda_1^{n-1}+q(n-1)\lambda_1^{n-2}\}\mathbf{u}^1+<\mathbf{x},\mathbf{v}^2>\underbrace{\{\lambda_1^{n+1}+p\lambda_1^n+q\lambda_1^{n-1}\}}_{=0}\mathbf{u}^2+\dots$$

$$=<\mathbf{x},\mathbf{v}^2>\{n\lambda^{n-2}\underbrace{(\lambda^2+p\lambda+q)}_{0}+\lambda^n-q\lambda^{n-2}\}\mathbf{u}^1+\dots=<\mathbf{x},\mathbf{v}^1>\lambda^{n-2}\underbrace{(\lambda^2-q)}_{0}\mathbf{u}^1+\dots\ ,$$

поскольку $(\lambda^2-q)=\frac{p^2}{4}-q=0$. Таким образом, $\mathbf{x}^{n+1}+p\mathbf{x}^n+q\mathbf{x}^{n-1}=o(\mathbf{x}^{n+1})$. При этом, координаты n-ой итерации $\mathbf{x}^k$ ведут себя как соответствующая степень λ : $x_i^n=(A^n\mathbf{x})_i\sim\lambda^n x_i$, поэтому естественно ввести три вектора $\mathbf{y}^{n+1,n,n-1}=\frac{\mathbf{x}^{n+1,n,n-1}}{\lambda^{n+1}}$. Для координат этих векторов, как следует из предыдущего, выполнено

$$y_k^{n+1}+py_k^n+qy_k^{n-1}=O(|\lambda'/\lambda|^{n+1})\ .$$

Выпишем соответствующие равенства для пары координат, скажем k и l.

$$\begin{array}{c|c|c} y_l^n & y_k^{n+1}+py_k^n+qy_k^{n-1}=O\left(|\lambda'/\lambda|^{n+1}\right)\sim 0 & y_l^{n-1} \\ y_k^n & y_l^{n+1}+py_l^n+qy_l^{n-1}=O\left(|\lambda'/\lambda|^{n+1}\right)\sim 0 & y_k^{n-1} \end{array}\ .$$

Домножая первое равенство на y_l^{n-1} , а второе на y_k^{n-1} и вычитая из первого равенства второе, получаем

$$p=-\frac{y_k^{n+1}y_l^{n-1}-y_l^{n+1}y_k^{n-1}}{y_k^ny_l^{n-1}-y_l^ny_k^{n-1}}+O\left(|\lambda'/\lambda|^{n+1}\right)=$$

$$=-\frac{x_k^{n+1}x_l^{n-1}-x_l^{n+1}x_k^{n-1}}{x_k^nx_l^{n-1}-x_l^nx_k^{n-1}}+O\left(|\lambda'/\lambda|^{n+1}\right)\ .$$

Аналогично, домножая первое равенство на y_l^n , а второе на y_k^n и вычитая их первого равенства второе, получаем

$$q=-\frac{x_k^{n+1}x_l^n-x_l^{n+1}x_k^n}{x_k^{n-1}x_l^n-x_l^{n-1}x_k^n}+O\left(|\lambda'/\lambda|^{n+1}\right)\ .$$

Заметим, что необходимое количество итераций в предложенном методе, можно контролировать исходя из того, что должно выполняться равенство $p^2/4=q$.

Глава 3

Обыкновенные дифференциальные уравнения

3.1 Общие сведения

Уравнение

$$F(x, u, u', \ldots, u^{(n)}) = 0$$

называется *обыкновенным дифференциальным уравнением n-го порядка*, если F определена и непрерывна в некоторой области $G \in \mathbf{R}^{n+2}$ $(n \geq 1)$ и, во всяком случае, зависит от $u^{(n)}$. Его решением является любая функция $u(x)$, которая этому уравнению удовлетворяет при всех x в определенном конечном или бесконечном интервале. Дифференциальное уравнение, разрешенное относительно старшей производной имеет вид

$$u^{(n)} = f(x, u, \ldots, u^{(n-1)}) \ . \tag{1}$$

Решением этого уравнения на интервале $I = [a, b]$ называется функция $u(x)$, такая что

1) $u(x) \in C^n[a, b]$,

2) $(x, u(x), \ldots, u^{(n-1)}(x)) \in D(f) \ \forall x \in I$,

3) $u^{(n)}(x) = f(x, u(x), \ldots, u^{n-1}(x)) \ \forall x \in I$.

3.1.1 Задача Коши

Задачей Коши (*начальной задачей*) для уравнения (1) называется задача нахождения такого решения уравнения (1), которое удовлетворяет начальным условиям

$$u(x_0) = u_0 \ , \ u'(x_0) = u_0' \ , \ldots \ , u^{(n-1)}(x_0) = u_0^{(n-1)} \ ,$$

где $u_0^{(i)}$ — некоторые заданные числа. Справедлива

Теорема Пеано. *Если f непрерывна в D, тогда для любой точки $x_0, u_0, \ldots, u_0^{(n-1)}$, принадлежащей области D, существует решение уравнения (1), определенное в некоторой окрестности точки $x_0 \in I$.*

Замечание. Теорема Пеано не гарантирует единственности.

Теорема Коши-Пикара. *Если f непрерывна в D и удовлетворяет условию Липшица по переменным u, u', ..., $u^{(n-1)}$, то есть*

$$|f(x;\mu_1,\mu_2,\dots,\mu_n) - f(x;\nu_1,\nu_2,\dots,\nu_n)| < L\sum_{k=1}^{n}|\mu_k-\nu_k| \ ,$$

то для любой точки $(x_0,u_0,\dots,u_0^{(n-1)})\in D$ существует единственное решение (1), определенное в некоторой окрестности точки $x_0\in I$.

Любое уравнение типа (1) можно свести к равносильной ему системе

$$\frac{du_i}{dx} = f_i(x;u_0,u_1\dots,u_{n-1}) \ , \quad i=0,1,,\dots,n-1 \ ,$$

дифференциальных уравнений первого порядка путем замены высших производных неизвестными функциями ($u_i(x) = u^{(i)}(x)$).

Теорему Коши-Пикара несложно доказать воспользовавшись теоремой о неподвижной точке сжимающего отображения [15]. Действительно, уравнение первого порядка

$$\begin{cases} u' = f(x,u) \\ u(x_0) = u_0 \end{cases}$$

эквивалентно интегральному уравнению

$$u(x) = u_0 + \int\limits_{x_0}^{x} f(t,u(t))dt \ .$$

По условию f непрерывна и, следовательно, $|f(x,u)|\le M$ в некоторой области $D'\subset D$, содержащей точку (x_0,u_0). Выберем $\delta>0$ так, чтобы:

1) $(x,u)\in D'$, если $|x-x_0|\le\delta$ и $|u-u_0|\le\delta M$;

2) $\delta L<1$, где L – константа, фигурирующая в условии Липшица.

Пусть C' – пространство всех непрерывных функций u, определенных при $|x-x_0|\le\delta$ и таких, что $|u(x)-u_0|\le\delta M$ с естественной для непрерывных функций метрикой $\rho(u_1,u_2) = \max\limits_x|u_1(x)-u_2(x)|$. Как замкнутое подпространство полного пространства $C_{[x_0-\delta,x_0+\delta]}$, пространство C' является полным. Убедимся, что отображение $y = Au$, определяемое формулой

$$y(x) = u_0 + \int\limits_{x_0}^{x} f(t,u(t))dt \ ,$$

является сжатием в C'. Действительно, пусть $u\in C'$ и $|x-x_0|\le\delta$, тогда

$$|y(x)-u_0| = \left|\int\limits_{x_0}^{x} f(t,u(t)dt\right| \le \delta M$$

и, следовательно, A переводит C' в себя. Далее,

$$|y_1(x)-y_2(x)| \le \int\limits_{x_0}^{x} |f(t,u_1(t)-f(t,u_2(t)|dt \le L\delta||u_1-u_2||_{C'} \ ,$$

и поскольку $\delta L<1$, то A – сжатие и, следовательно, в C' существует единственное решение уравнения $u = Au$. Аналогично доказывается однозначная разрешимость задачи Коши для системы уравнений первого порядка, а, следовательно, и для задачи Коши произвольного порядка.

3.1.2 Краевая задача

Сформулируем *краевую задачу* только для уравнений второго порядка, являющуюся одной из самых существенных. Такая задача имеет вид:

$$\begin{cases} u'' = f(x,u,u'), \;\; x \in [a,b], \\ \alpha_1 u(a) + \beta_1 u'(a) = \gamma_1, \\ \alpha_2 u(b) + \beta_2 u'(b) = \gamma_2, \end{cases}$$

где в краевых условиях считается, что $|\alpha_i| + |\beta_i| \neq 0$, $i = 1,2$. В отличие от задачи Коши здесь значительно сложнее исследуется вопрос о существовании решения. Очень важный и наиболее часто встречающийся случай — линейное дифференциальное уравнение второго порядка

$$u'' + p(x)u' + q(x)u = f(x) \;,$$

краевую задачу для которого мы и будем рассматривать в дальнейшем.

3.1.3 Задача Штурма-Лиувилля

Задача Штурма-Лиувилля или задача на собственные функции и собственные значения является одновременно и краевой задачей (с однородными краевыми условиями) и обычно записывается в, так называемом, самосопряженном виде:

$$-\frac{d}{dx}\left[k(x)\frac{du}{dx}\right] + [q(x) - \lambda r(x)]\, u(x) = 0 \;,$$

$$\alpha_1 u(a) + \beta_1 u'(a) = 0 \;, \quad \alpha_2 u(b) + \beta_2 u'(b) = 0 \;.$$

Здесь требуется найти те λ, при которых задача разрешима (собственные значения), и соответствующие им решения $u_\lambda(x)$ — собственные функции, определяемые с точностью до постоянного множителя.

3.1.4 Что понимается под численным решением

Точные (аналитические) методы решения — такие методы, когда решение дифференциального уравнения можно получить в виде элементарных функций или квадратур от них, что, естественно, возможно не всегда. Численные методы — методы нахождения решений не на всем промежутке изменения независимой переменной, а лишь в дискретном наборе точек $x_0, x_1, \ldots, x_N \in [a,b]$. Здесь, правда, следует отметить, что можно искать решение в виде разложения в ряд по некоторой полной системе функций (скажем, в ряд Фурье) и обрезать его на некотором члене. Однако, вопрос о том, какую систему функций использовать и какое количество членов разложения использовать, является одновременно и численным и аналитическим. Численные методы применимы к очень широкому классу дифференциальных уравнений. В соответствии с двумя типами задач для дифференциальных уравнений, численные методы тоже делятся на два класса: Численные методы решения задачи Коши и численные методы решения краевой задачи и задачи Штурма-Лиувилля.

3.2 Задача Коши

Рассмотрим задачу Коши для уравнения первого порядка на отрезке $[a,b]$:

$$u' = f(x,u) \;, \quad u(a) = u_0 \;, \qquad (2)$$

Разобьём промежуток $[a, b]$ на N частей $a = x_0 < x_1, < \ldots, < x_N$. Обозначим $u(x_i) = u_i$, где $u(x)$ — точное решение задачи Коши, и через y_i значения приближенного решения в точках x_i . Существует два типа численных схем :

1. явные : $y_i = F(y_{i-k}, y_{i-k+1}, \ldots, y_{i-1})$ (а);
2. неявные : $y_i = F(y_{i-k}, y_{i-k+1}, \ldots, y_i)$ (б).

Здесь F некоторая функция, связывающая приближения. В явных схемах приближенное значение y_i в точке x_i определяется через некоторое число k уже определенных приближенных значений. В неявных схемах y_i определяется не рекуррентным образом как в явных схемах, а для его определения возникает уравнение, поскольку равенство (б) представляет из себя именно уравнение на y_i. Явные схемы проще, однако зачастую неявные схемы предпочтительнее.

3.2.1 Получение явных схем

Обширный класс явных схем для решения задачи Коши получается с помощью разложения в ряд Тейлора. Выпишем его для функции $u(x)$:

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \ldots + \frac{h^n}{n!}u^{(n)}(x) + \ldots .$$

Если $u(x)$ — решение задачи (2), то $u'(x_i) = f(x_i, u_i)$ и, следовательно $u''(x_i) = \frac{d}{dx}f(x,u)|_{x_i} = f'_x(x_i, u_i) + f(x_i, u_i)f'_u(x_i, u_i)$. Поступая далее таким же образом, можно выразить все производные $u^{(k)}$ через производные известной функции $f(x, u)$:

$$u_{i+1} = u_i + hf(x_i, u_i) + \frac{h^2}{2}[f'_x(x_i, u_i) + f(x_i, u_i)f'_u(x_i, u_i)] + \ldots . \tag{3}$$

Обрывая (3) на том или ином члене, получаем различные явные схемы для вычисления приближенного решения с определенной степенью точности по h.

3.2.2 Схема Эйлера (метод ломаных)

Оставляя в (3) только члены первого порядка по h, получаем приближенное равенство: $u_{i+1} \approx u_i + hf(x_i, u_i)$. Заменяя в нем точные значения $u_i = u(x_i)$ на приближения y_i, получаем приближенную схему:

$$\begin{cases} y_0 = u_0 \\ y_{i+1} = y_i + hf(x_i, y_i) \end{cases}, \quad i = 0, 1, \ldots, N .$$

Указанная процедура является *методом Эйлера* и имеет первый порядок сходимости по h , если $f(x, u)$ ограничена и ограничены ее первые производные по обоим аргументам. Убедимся в этом. Действительно, пусть $c = \max\limits_{x,u}\{|f|, |f'_x|, |f'_u|\}$. Обозначим разность между истинным решением u_j в точке x_j и найденным по методу Эйлера приближением y_j через v_j , тогда

$$v_{j+1} = v_j + h\underbrace{[f(x_j, u_j) - f(x_j, y_j)]}_{f'_y(x_j, \tilde{y}_j)v_j} + \frac{h^2}{2}u''(x_j) + O(h^3) ,$$

где $\tilde{y}'_j$ — некоторая точка между u_j и y_j . Заметим, что поскольку $y_0 = u_0$, то $v_0 = 0$. Тогда $v_1 = \frac{1}{2}h^2u''_0 + O(h^3)$, и далее

$$v_2 = v_1(1 + hf'_u(x_1, \tilde{y}'_1)) + \frac{1}{2}h^2u''_1 + (h^3)$$

$$\frac{1}{2}h^2\left(u_1'' + u_0''\left[1 + hf_u'(x_1,\tilde{y}_1)\right]\right) + O(h^3)\ ,$$

$$\vdots$$

$$v_{j+1} = \frac{1}{2}h^2\sum_{k=0}^{j}u_k''\prod_{i=k+1}^{j}\left[1 + hf_u'(x_i,\tilde{y}_i)\right] + O(h^3) =$$

$$= \frac{1}{2}h^2\sum_{k=0}^{j}u_k''\underbrace{\left[1 + \underbrace{\sum_{i=k+1}^{j}hf_u'(x_i,\tilde{y}_i)}_{<c(x_j-x_{k+1})}\right]}_{<\exp\{c(x_j-x_{k+1})\}} + O(h^3)\ .$$

Поскольку $u'' = f_x' + ff_u'$, то $|u''| \le c + cc \equiv c_1$, и

$$|v_{j+1}| \le \frac{1}{2}hc_1\sum_{k=0}^{j}he^{c(x_j-x_k)} = \frac{1}{2}hc_1\int\limits_{x_0}^{x_j}e^{c(x_j-t)}dt + o(h) =$$

$$= h\frac{c_1}{2c}\left[e^{c(x_j-x_0)} - 1\right] + o(h) = O(h)\ .$$

Таким образом, метод Эйлера имеет первый порядок точности по h и при достаточно малом шаге приближенное решение близко к точному.

3.2.3 Методы Рунге-Кутта

Метод Рунге-Кутта 2-го порядка

Выпишем ряд Тейлора для решения дифференциального уравнения $u(x)$ с точностью до квадратичных членов

$$u_{j+1} = u_j + hf(x_j,u_j) + \frac{h^2}{2}\underbrace{\left[f_x'(x_j,u_j) + f(x_j,u_j)f_u'(x_j,u_j)\right]}_{u''(x_j)} + \ldots\ . \tag{4}$$

Сама по себе такая схема уже годится для приближенного решения дифференциального уравнения, однако ее неудобство состоит в том, что приходится дифференцировать функцию $f(x,u)$ по обоим аргументам. Если заменить эти производные разностными, то формально можно записать

$$u_{j+1} = u_j + h\left[\alpha f(x_j,u_j) + \beta f(x_j + \gamma h, u_j + \delta h)\right] + \ldots\ , \tag{5}$$

где константы $\alpha,\beta,\gamma,\delta$ необходимо определить исходя из того, что эти два представления должны совпадать с точностью до $O(h^3)$. Для этого разложим в (5) $f(x_j + \gamma h, u_j + \delta h)$ в ряд Тейлора:

$$u_{j+1} = u_j + h(\alpha + \beta)f(x_j,u_j) + \beta h^2\left[\gamma f_x'(x_j,u_j) + \delta f_u'(x_j,u_j)\right] + O(h^3)\ .$$

Сравнивая с (4), получаем 3 уравнения на 4 неизвестных коэффициента: $\alpha + \beta = 1$, $\beta\gamma = \frac{1}{2}$, $\beta\delta = \frac{1}{2}f(x_j,u_j)$. Выразив их через β и заменив истинные значения $u_j = u(x_j)$ на приближенные y_j и отбросив кубические члены, получаем набор разностных схем Рунге-Кутта 2-го порядка

$$y_{j+1} = y_j + h\left[(1-\beta)f(x_j,y_j) + \beta f(x_j + \frac{h}{2\beta}, y_j + \frac{h}{2\beta}f(x_j,y_j))\right]\ ,\quad 0 < \beta \le 1\ .$$

Обычно полагают β равным 1/2 или 1.

Метод Рунге-Кутта 4-го порядка

Изложенным выше способом можно строить схемы типа Рунге-Кутта различного порядка точности по h . В частности, метод Эйлера является схемой Рунге-Кутта 1-го порядка. Наиболее удобной и употребительной является схема 4-го порядка. Она имеет следующий вид

$$y_{j+1} = y_j + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) ,$$

$$k_1 = f(x_j, y_j) , \quad k_2 = f\left(x_j + h/2, y_j + hk_1/2\right) ,$$

$$k_3 = f\left(x_j + h/2, y_j + hk_2/2\right) , \quad k_4 = f(x_j + h, y_j + hk_3) .$$

На каждом шаге величины k_m рассчитываются заново.

Интересно отметить, что если f есть функция только от x , то решение уравнения есть $u(x) = u_0 + \int\limits_{x_0}^{x} f(t)dt$, и формулы Рунге-Кутта превращаются в формулы приближенного интегрирования. Методу Эйлера соответствует формула левых прямоугольников, методу Рунге-Кутта 2-го порядка с $\beta = 1$ соответствует формула средних, а с $\beta = 1/2$ — формула трапеций. Наконец, методу Рунге-Кутта 4-го порядка соответствует формула Симпсона с шагом $h/2$. Это косвенно свидетельствует о порядке точности той или иной схемы.

Естественным образом схемы Рунге-Кутта обобщаются на случай систем уравнений 1-го порядка при помощи формальной замены функций $y(x)$ и $f(x,y)$ на вектор-функции $\mathbf{y}(x)$ и $\mathbf{f}(x,\mathbf{y})$. При этом, поскольку уравнение n-го порядка эквивалентно системе из n уравнений 1-го порядка, то методы Рунге-Кутта можно применять к задаче Коши для уравнений порядка выше 1-го. В частности, рассмотрим задачу Коши для уравнения 2-го порядка

$$\begin{cases} u'' = f(x,u,u') \\ u(x_0) = u_0 \\ u'(x_0) = u'_0 \end{cases} .$$

Обозначим $u' = v$ и введем вектор $\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}$, тогда система принимает вид

$$\begin{cases} \mathbf{u} = \mathbf{f}(x,\mathbf{u}) \\ \mathbf{u}(x_0) = \mathbf{u}_0 \end{cases} \qquad \begin{cases} \frac{d}{dx}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ f(x,u,v) \end{pmatrix} \\ \begin{pmatrix} u(x_0) \\ v(x_0) \end{pmatrix} = \begin{pmatrix} u_0 \\ u'_0 \end{pmatrix} \end{cases} .$$

Если ввести вектор $\mathbf{y}_j = \begin{pmatrix} y_j \\ z_j \end{pmatrix}$ приближений к истинному решению $\mathbf{u}_j$ в точке x_j и вектора $\mathbf{k}_m = \begin{pmatrix} k_m \\ q_m \end{pmatrix}$ расчетных коэффициентов, то метод Рунге-Кутта 4-го порядка принимает вид

$$\mathbf{y}^{(j+1)} = \begin{pmatrix} y_{j+1} \\ z_{j+1} \end{pmatrix} = \begin{pmatrix} y_j + h(k_1 + 2k_2 + 2k_3 + k_4)/6 \\ z_j + h(q_1 + 2q_2 + 2q_3 + q_4)/6 \end{pmatrix}$$

$$k_1 = z_j , \quad k_2 = z_j + \frac{h}{2}q_1 , \quad k_3 = z_j + \frac{h}{2}q_2 , \quad k_4 = z_j + hq_3 ,$$

$$q_1 = f(x_j, y_j, z_j) , \quad q_2 = f(x_j + \frac{h}{2}, y_j + \frac{h}{2}k_1, z_j + \frac{h}{2}q_1) ,$$

$$q_3 = f(x_j + \frac{h}{2}, y_j + \frac{h}{2}k_2, z_j + \frac{h}{2}q_2), \; q_4 = f(x_j + h, y_j + hk_3, z_j + hq_3) .$$

3.2.4 Методы Адамса

Явная схема Адамса

Рассмотренные выше схемы являются явными одношаговыми (для нахождения последующего приближения используется лишь одно предыдущее). Приводимые ниже методы являются многошаговыми. Они могут быть как явными, так и неявными.

Пусть задана задача Коши

$$\begin{cases} u' = f(x,u), \\ u(a) = u_0. \end{cases}$$

Для точного решения $u(x)$ (которое нам неизвестно) выполнено

$$u(x_{n+1}) = u(x_n) + \int\limits_{x_n}^{x_{n+1}} f(x, u(x))dx \ . \qquad (6)$$

Предположим, нам известны приближенные значения y_i функции $u(x)$ в k точках $x_{n-k+1}, x_{n-k+2}, \ldots, x_n$ (стартовые k точек, в частности, можно найти методом Эйлера или методами Рунге-Кутта того или иного порядка), тогда функцию $f(x, u(x))$ в (6) для приближенного вычисления интеграла можно заменить на интерполяционный полином $P_{n,k}(x)$ порядка $k-1$, построенный по k точкам $\{x_i, f(x_i, y_i)\}_{n-k+1}^{n}$, интеграл от которого считается явно и представляет собой линейную комбинацию значений $f_i = f(x_i, y_i)$ с некоторыми множителями λ_i. Таким образом, мы получаем следующую рекуррентную процедуру вычисления приближенных значений y_i функции $u(x)$ (являющейся точным решением задачи Коши) в точках x_i

$$y_{n+1} = y_n + \int\limits_{x_n}^{x_{n+1}} P_{n,k}(x)dx = y_n + \sum_{i=1}^{k} \lambda_i f(x_{n+1-i}, y_{n+1-i}) \ . \qquad (7)$$

Описанная схема называется k-шаговой явной формулой Адамса.

Неявная схема Адамса. Метод прогноз-коррекции

Пусть $P_{n+1,k+1}(x)$ — интерполяционный полином порядка k, построенный по $k+1$ значению $f_{n-k+1}, \ldots, f_n, f_{n+1}$, одно из которых, именно f_{n+1}, мы будем считать неизвестным. Модифицируем (7), заменив в нем $P_{n,k}$ на полином более высокой степени $P_{n+1,k+1}$, интеграл от которого выражается в виде линейной комбинации значений f_i с некоторыми новыми коэффициентами β_i :

$$y_{n+1} = y_n + \int\limits_{x_n}^{x_{n+1}} P_{n+1,k+1}dx = y_n + \sum_{i=1}^{k} \beta_i f_{n+1-i} + \beta_0 f(x_{n+1}, y_{n+1}) \ . \qquad (8)$$

Формула (8) представляет собой неявную схему Адамса и является уравнением на y_{n+1}, которое можно решать скажем методом последовательных приближений. Естественно, что начальное приближение y^0_{n+1}, должно быть разумно выбрано. Для этого удобно объединить явную и неявную схемы Адамса в одну, называемую методом "прогноз-коррекции". Именно, с помощью явной схемы определяется начальное приближение y^0_{n+1} (прогноз), а затем по неявной схеме оно необходимое число раз (обычно один или два) корректируется методом последовательных приближений до достижения заданной точности (коррекция):

прогноз: $\quad y^0_{n+1} = y_n + \sum\limits_{i=1}^{k} \lambda_i f_{n+1-i},$

коррекция: $\quad y^{m+1}_{n+1} = y_n + \sum\limits_{i=1}^{k} \beta_i f_{n+1-i} + \beta_0 f(x_{n+1}, y^m_{n+1}).$

Пример. Пусть k равно 1 и $h = x_{n+1} - x_n$. В этом случае "прогноз" представляет собой интегрирование по формуле левых прямоугольников, совпадающее в данном случае с методом Эйлера, а "коррекция" интегрирование по формуле трапеций:

прогноз: $y^0_{n+1} = y_n + hf_n$,

коррекция: $y_{n+1} = y_n + \frac{h}{2}(f_n + f_{n+1})$.

Последнюю формулу необходимо понимать как уравнение для определения y_{n+1} (и, соответственно, уравнение на $f_{n+1} = f(x_{n+1}, y_{n+1})$), которое решается методом последовательных приближений.

3.3 Краевая задача

3.3.1 Метод стрельбы

Рассмотрим краевую задачу для уравнения второго порядка

$$\begin{cases} y''(x) = f(x, y, y') \ , \quad x \in [a, b] \ , \\ \alpha_1 y(a) + \beta_1 y'(a) = \gamma_1 \ , \\ \alpha_2 y(b) + \beta_2 y'(b) = \gamma_2 \ . \end{cases} \tag{9}$$

Перейдем от этой задачи к системе уравнений первого порядка. Пусть $u(x) = y(x)$ и $v(x) = y'(x)$. Тогда уравнение (9) переходит в

$$\begin{cases} u' = v, \\ v' = f(x, u, v), \end{cases} \tag{10}$$

а краевые условия принимают вид

$$\begin{cases} \alpha_1 u(a) + \beta_1 v(a) = \gamma_1, \\ \alpha_2 u(b) + \beta_2 v(b) = \gamma_2. \end{cases} \tag{10'}$$

Таким образом, исходная краевая задача свелась к задаче 1-го порядка для системы двух уравнений.

Метод стрельбы — это переход к решению некоторой задачи Коши для системы (10). Выберем произвольно $u(a) = \xi$. Теперь определим $v(a)$ из первого из условий (10′):

$$v(a) = \beta_1^{-1}(\gamma_1 - \alpha_1 \xi) \equiv \eta(\xi) \ .$$

Далее, рассмотрим систему (10) с начальными условиями

$$\begin{cases} u(a) = \xi \\ v(a) = \eta(\xi) \ . \end{cases}$$

Такая задача является задачей Коши. Решим ее некоторым способом (например, методом Рунге-Кутта 4-го порядка). Решение (u_ξ, v_ξ) наверняка не будет удовлетворять второму краевому условию. Обозначим через Δ_ξ возникающую невязку:

$$\alpha_2 u(b)_\xi + \beta_2 v(b)_\xi - \gamma_2 = \Delta(\xi) \ .$$

Задача состоит в отыскании такого ξ_*, при котором невязка обращается в ноль: $\Delta(\xi_*) = 0$, что соответствует удовлетворению второго краевого условия. Варьируем (стрельба) пристрелочный параметр ξ до тех пор, пока не образуется вилка

$$\xi_i \ : \ \Delta(\xi_i)\Delta(\xi_{i+1}) < 0 \ ,$$

тогда можно утверждать, что $\xi_* \in [\xi_i, \xi_{i+1}]$. После того, как промежуток на котором находится корень функции $\Delta(\xi)$ найден, делим отрезок $[\xi_i, \xi_{i+1}]$ пополам и выбираем ту его часть, на концах которой Δ имеет разные знаки, и так далее, до достижения требуемой точности.

Замечание. при каждом выбранном ξ_i необходимо решать задачу Коши дифференциального уравнения (10) с начальными данными

$$u(a) = \xi_i \ , \ \ v(a) = \eta(\xi_i) \ .$$

3.3.2 Метод сеток (разностный метод)

Рассмотрим разностный метод на примере следующего дифференциального уравнения второго порядка:

$$\begin{cases} -u'' + q(x)u = f(x) \quad [a,b], \\ u(a) = A \ , \ u(b) = B \ . \end{cases} \tag{11}$$

Разобьем промежуток на N частей: $a = x_0 < x_1 < \ldots < x_N = b$. Пусть шаг сетки постоянный: $x_i - x_{i-1} = h$. Аппроксимируем вторую производную $u''(x_i)$ разностной:

$$u''(x_i) = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} - \frac{u^{(4)}(x_i)h^2}{12} + O(h^4) \ ,$$

выражение для которой легко получить из ряда Тейлора

$$u(x_i \pm h) = u(x_i) \pm u'(x_i)h + \frac{u''(x_i)h^2}{2} \pm \frac{u'''(x_i)h^3}{6} + \frac{u^{(4)}(x_i)h^4}{24} + \ldots \ ,$$

Введем обозначения: $u(x_i) = u_i$, $q_i = q(x_i)$, $f_i = f(x_i)$. Заменим в (11) вторую производную разностной, тогда для приближенного решения y_i в точках x_i получаем трехдиагональную систему

$$-y_{i-1} + (2 + h^2 q_i)y_i - y_{i+1} = f_i h^2 \ , \ i = 1, 2, \ldots, N-1 \ . \tag{12}$$

Для ее разрешимости достаточным условием (но вовсе не необходимым) является диагональное преобладание. В нашем случае это сводится к требованию $|2 + h^2 q_i| > 2$, которое выполняется если $q(x) > 0$.

3.3.3 Сходимость сеточных методов

Пусть $u(x)$ — точное решение уравнения (11), а y_i — численное решение задачи (12). Справедлива

Теорема. *Пусть* $q(x),\ f(x) \in C^2_{[a,b]}$ *и* $q(x) > 0$, $\forall\ x \in [a,b]$, *тогда*

$$|u(x_i) - y_i| = O(h^2) \ .$$

Доказательство. Поскольку $q(x), f(x) \in C^2_{[a,b]}$ то из уравнения (11) следует, что $u(x) \in C^4[a,b]$, и тогда используя ряд Тейлора можно записать

$$u''(x_i) = \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} - \frac{1}{12}h^2 u^{(4)}(\xi_i) \ , \ \xi_i \in (x_{i-1}, x_i) \ .$$

Значения u_i точного решения удовлетворяет уравнениям

$$-\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + qu_i = f_i - \frac{1}{12}h^2 u^{(4)}(\xi_i) \ ,$$

где ξ_i некоторые точки на $[a,b]$. Для погрешности

$$v_i = y_i - u_i$$

возникает система уравнений

$$-\frac{v_{i-1} - 2v_i + v_{i+1}}{h^2} + q_i v_i = \frac{1}{12}h^2 u^{(4)}(\xi_i) \quad , \ \ v_0 = 0 \ , \ v_N = 0 \ . \tag{13}$$

Пусть x_k — точка, где модуль погрешности максимален, то есть

$$|v_k| \geq |v_i| \ , \ i = 1, 2, \dots, N-1 \ .$$

Этой точкой не может являться x_0 и x_N, поскольку $v_0 = v_N = 0$. Сравним модули левой и правой части системы (13) при индексе равном k

$$|v_k(2 + q_k h^2)| \leq |v_{k-1}| + |v_{k+1}| + \frac{1}{12} h^4 |u^{(4)}(\xi_k)| \ ,$$

или

$$|v_k|(2 + q_k h^2) \leq 2|v_k| + \frac{1}{12} h^4 |u^{(4)}(\xi_k)| \ ,$$

откуда

$$|v_k| \leq \frac{1}{12} h^2 \frac{|u^{(4)}(\xi_k)|}{|q_k|} \ ,$$

то есть

$$\max_i |v_i| \leq \frac{h^2}{12} \max_i \frac{|u^{(4)}(\xi_i)|}{|q_i|} \ ,$$

что и требовалось доказать.

3.3.4 Метод Нумерова

Точность сеточного метода (12) можно повысить до четвертого порядка несколько модифицировав его *методом Нумерова*, справедливым для более широкого класса уравнений. Именно, для уравнений вида

$$u'' = f(x, u) \ . \tag{14}$$

Подставим в (14) вместо второй производной разностную:

$$0 = u''(x) - f(x,u) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} - f(x) - \frac{h^2 u^{(4)}(x)}{12} + O(h^4) \ . \tag{15}$$

Непосредственно из уравнения (14) следует, что $u^{(4)} = f''(x,u)$. Заменим в (15) четвертую производную от неизвестной функции в точке x_i на вторую от $f(x,u)$, которую в свою очередь заменим разностной

$$f(x,u)''_i = \frac{f(x_{i+1}, u_{i+1}) + f(x_{i-1}, u_{i-1}) - 2f(x_i, u_i)}{h^2} + O(h^2) \ .$$

Тот факт, что точность такой формулы действительно имеет второй порядок, необходимо еще проверять. Здесь мы не будем останавливаться на этом (подробнее см. [2]). Имеем

$$u''_i - f(x_i, u_i) \\ = \frac{u_{i+i} + u_{i-1} - 2u_i}{h^2} - f(x_i, u_i) - \frac{h^2}{12} \left[\frac{f(x_{i+1}, u_{i+1}) + f(x_{i-1}, u_{i-1}) - 2f(x_i, u_i)}{h^2} + O(h^2) \right] \ ,$$

то есть численная схема приобретает вид

$$\frac{y_{i+i} + y_{i-1} - 2y_i}{h^2} = \frac{1}{12} \left[f(x_{i+1}, y_{i+1}) + f(x_{i-1}, y_{i-1}) + 10 f(x_i, y_i) \right] .$$

В частности, для уравнения (11)

$$u_{i+1}\left(1 - \frac{q_{i+1} h^2}{12}\right) - u_i\left(2 + h^2 q_i \frac{5}{6}\right) + u_{i-1}\left(1 - \frac{q_{i-1} h^2}{12}\right) \\ = -\frac{h^2}{12}(f_{i+1} + f_{i-1} + 10 f_i) + O(h^6) \ .$$

Отбрасывая остаточный член и добавляя граничные условия в точках x_0 и x_N, получаем сеточный метод с погрешностью $0(h^4)$ (напомним, что в обычном методе сеток было:
$u_{i+1} - u_i(2 + h^2 q_i) + u_{i-1} = -f_i h^2 + O(h^4)$.)

3.4 Задача Штурма-Лиувилля

Задачу на собственные значения рассмотрим на примере следующего дифференциального уравнения 2-го порядка:

$$\begin{cases} -u'' + q(x)u = \lambda u, \\ u(a) = 0 \ , \ u(b) = 0. \end{cases} \tag{16}$$

Вопрос. Почему граничные условия однородные (нулевые)?

В задаче появилась новая степень свободы — λ. Важные свойства задачи (16) таковы, что решение дифференциального уравнения существует и удовлетворяет граничным условиям лишь при некоторых значениях λ , называемых собственными значениями. Соответствующие этим λ решения $u_\lambda(x)$ называются собственными функциями. Спектр собственных значений может быть дискретным (в рассматриваемом случае спектр дискретен, если и a и b конечны), непрерывным, также λ может одновременно принадлежать дискретному и непрерывному спектру. В задаче (16) требуется определить как возможные значения λ, так и собственные функции $u_\lambda(x)$.

Существует 2 основных метода решения задачи (16).

3.4.1 Метод стрельбы

В силу однородности задачи (16), если $u(x)$ является решением, то $u_1(x) = \text{const}\, u(x)$ тоже решение, поэтому можно задать произвольно значение $u'(x)$ в точке a (обычно выбирают $u'(a) = 1$), а затем перейти к стрельбе, то есть рассмотреть задачу Коши:

$$\begin{cases} -u'' + q(x)u = \lambda u \\ u(a) = 0 \\ u'(a) = 1 \end{cases}$$

и находить ее решение $u(x, \lambda)$ и подобрать λ так, чтобы

$$u(b, \lambda) = 0 \ . \tag{17}$$

При этом мы одновременно находим и собственное значение λ и соответствующую собственную функцию $u(x, \lambda)$. Решается уравнение (17) любым из методов нахождения корня алгебраического уравнения. Например, варьируя пристрелочный параметр, можно добиться вилки $u(b, \lambda_i)u(b, \lambda_{i+1}) < 0$ и затем использовать метод деления пополам.

Метод стрельбы удобно применять в ситуации, когда априори из физической постановки задачи известны естественные пристрелочные параметры.

3.4.2 Метод сеток

Разобьем промежуток на N частей введя сетку $a = x_0 < x_1 < \ldots < x_N = b$, и также, как в случае краевых задач, заменим в (16) производные разностными. При этом задача принимает вид

$$\begin{cases} y_{i-1} - (2 + h^2 q_i) y_i + y_{i+1} = \lambda h^2 y_i \ , \\ y_0 = 0 \ , \\ y_N = 0 \ . \end{cases}$$

Таким образом, исходная задача свелась к задаче на собственные значения для трехдиагональной матрицы A размера $(N-1)\times(N-1)$:

$$Ay = \lambda y \ ,$$

$$A \ : \quad \begin{array}{l} a_{ii} = 2 + h^2 q_i \\ a_{i-1\,i} = a_{i\,i+1} = -1 \end{array} \qquad i = 1, 2, \ldots, N-1 \ .$$

Собственные числа матрицы A являются приближениями к первым собственным значениям исходной задачи.

3.5 Разностный оператор второй производной

3.5.1 Оператор второй производной

Произведем сначала спектральный анализ собственно оператора второй производной на отрезке $[a, b]$ с нулевыми граничными условиями, т.е. определим его собственные функции и собственные числа.

$$\begin{cases} -\frac{d^2}{dx^2}\Phi = \lambda\Phi, \\ \Phi(a) = \Phi(b) = 0. \end{cases} \tag{18}$$

Очевидно, что функции $\Phi_\lambda(x) = e^{\pm i\sqrt{\lambda}x}$, или их комбинации $\sin\sqrt{\lambda}x$, $\cos\sqrt{\lambda}x$ удовлетворяют уравнению. Пусть $a = 0$ для упрощения записи. Поскольку $\Phi(0) = 0$, то нас устраивает только функции вида $\sin\sqrt{\lambda}x$. Из второго граничного условия $\Phi(b) = 0$ следует, что $\sqrt{\lambda}b = \pi n$, таким образом, спектр задачи дискретный и бесконечный. Собственные функции Φ^n и собственные числа λ_n имеют вид

$$\Phi^n(x) = \sin\frac{\pi n}{b}x \ , \quad \lambda_n = \frac{n^2\pi^2}{b^2} \ . \tag{19}$$

3.5.2 Разностный оператор

Рассмотрим теперь соответствующую разностную задачу. Разобьем промежуток на $(N+1)$ часть с равномерным шагом h : $a = x_0 < x_1 < \ldots < x_{N+1} = b$. Задача на спектр разностного оператора принимает вид

$$\begin{cases} -\frac{F_{i-1} - 2F_i + F_{i+1}}{h^2} = \tilde{\lambda}F_i \ , \quad i = 1, 2, \ldots, N, \\ F_0 = F_{N+1} = 0, \end{cases} \tag{20}$$

или, обозначив $\tilde{\lambda}h^2 = \mu$,

$$\begin{cases} -F_{i-1} + 2F_i - F_{i+1} = \mu F_i \ , \quad i = 1, \ldots, N, \\ F_0 = F_{N+1} = 0. \end{cases}$$

Эта задача представляет собой задачу на спектр трехдиагональной матрицы N-го порядка

$$A\mathbf{F} = \mu\mathbf{F} \ : \quad \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots \\ 0 & -1 & 2 & -1 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \ldots \\ \ldots \\ F_N \end{pmatrix} = \mu \begin{pmatrix} F_1 \\ F_2 \\ \ldots \\ \ldots \\ F_N \end{pmatrix} \ ,$$

$$F_0 = F_{N+1} = 0 \ ,$$

с N-компонентными собственными векторами $\mathbf{F} = (F_1, F_2 \dots, F_N)^T$.

Для решения этой задачи вспомним сначала (см. Главу "Численное дифференцирование), что $e^{h\frac{d}{dx}}F(x) = F(x+h)$, то есть $e^{\pm h\frac{d}{dx}}F_i = F_{i\pm1}$. Таким образом, систему можно переписать в виде

$$\begin{cases} [e^{-h\frac{d}{dx}} + 2 - e^{h\frac{d}{dx}}]F_i = \mu F_i, \\ F_0 = F_{N+1} = 0. \end{cases}$$

Применяя операторы сдвига ко всем компонентам вектора $\mathbf{F}$, получаем следующую переформулировку

$$\begin{cases} [e^{-h\frac{d}{dx}} + 2 - e^{h\frac{d}{dx}}]\mathbf{F} = \mu \mathbf{F}, \\ F_0 = F_{N+1} = 0. \end{cases}$$

Некоторое неудобство такой формы записи состоит в том, что $\frac{d}{dx}$ не является самосопряженным оператором, но таковым является оператор $\frac{1}{i}\frac{d}{dx}$ (правда рассматриваемый на всей оси):

$$\left\langle \frac{1}{i}\frac{d}{dx}f, g \right\rangle = \frac{1}{i}\int f'(x)g(x)dx = \int f(x)\left(\overline{\frac{1}{i}\frac{d}{dx}g(x)}\right)dx = \left\langle f, \frac{1}{i}\frac{d}{dx}g \right\rangle .$$

Собственные функции оператора D это экспоненты e^{ipx} : $\frac{1}{i}\frac{d}{dx}e^{ipx} = pe^{ipx}$, спектр сплошной и заполняет всю вещественную ось: $p \in \mathbf{R}^1$. Но собственные функции произвольного самосопряженного оператора A являются собственными и для функции от оператора $f(A)$, а собственные значения оператора $f(A)$ это числа $f(p)$), где p — собственные числа A :

$$\begin{matrix} A\varphi = \sum \lambda_i \langle \varphi, F^{(i)} \rangle F^{(i)} \\ f(A)\varphi = \sum f(\lambda_i) \langle \varphi, F^{(i)} \rangle F^{(i)} \end{matrix} \quad \Rightarrow \quad f(A)F^{(k)} = f(\lambda_k)F^{(k)} .$$

Подействуем на собственную функцию $F = e^{ipx}$ оператора дифференцирования D функцией $f(D) = [-e^{ihD} - e^{-ihD} + 2]$ от этого оператора:

$$[-e^{ihD} - e^{-ihD} + 2]F = [-e^{iph} - e^{-iph} + 2]F = 2[1 - \cos ph]F .$$

В силу симметрии $f(D)$ очевидно что $f(p) = f(-p)$, поэтому собственная функция e^{-ipx} отвечает тому же собственному числу $2[1 - \cos(ph)]$, что и e^{ipx} (равно как и любая их линейная комбинация). В нашей задаче необходимо удовлетворить граничным условиям $F(0) = F(a) = 0$. Из первого граничного условия $F_0 = 0$ следует, что компоненты собственного вектора, отвечающего собственному числу p , имеют вид $F_j^p = \sin px_j$, где $x_j = hj$. Второе граничное условие $F_{N+1} = 0$ позволяет определить сами собственные числа: $\sin ph(N+1) = 0$, откуда $ph(N+1) = \pi n$, или $p_n = \frac{\pi n}{h(N+1)} = \frac{\pi n}{b}$, $n = 1, 2, \dots, N$. То есть в задаче (20) собственные векторы имеют вид

$$\mathbf{F}^n \ : \ F_j^n = \sin\frac{\pi n}{b}x_j \ , \ x_j = hj .$$

Заметим, что значение истинной собственной функции Φ^n оператора двойного дифференцирования в любой точке x_j совпадает с j-компонентой n-го собственного вектора разностного оператора:

$$\Phi^n(x_j) = F_j^n .$$

Посмотрим теперь насколько отличаются собственные значения λ_n оператора двойного дифференцирования и собственные числа $\tilde{\lambda}_n = \frac{\mu_n}{h^2}$ разностного оператора (19):

$$\tilde{\lambda}_n = \frac{\mu_n}{h^2} = \frac{2}{h^2}[1 - \cos p_n h] = \frac{2}{h^2}[1 - \cos\frac{\pi n}{b}h] =$$

$$= \frac{2}{h^2}[1 - 1 + \frac{\pi^2 n^2}{2b^2}h^2 + O(h^4)] = \frac{\pi^2 n^2}{b^2} + O(h^2) = \lambda_n + O(h^2) .$$

3.5.3 Резольвента

Определение. Пусть A — линейный оператор, функция от оператора $R_\lambda(A) = (A-\lambda)^{-1}$ называется *резольвентой* оператора A .

Резольвента $R_\lambda(A)$ определена , как легко видеть, не при всех λ, а лишь вне спектра.

Пусть A — самосопряженный оператор с дискретным спектром, λ_k — его собственные числа, φ^k — соответствующие собственные функции. Выпишем спектральное разложение A :

$$A = \sum \lambda_k P_k = \sum \lambda^k \langle\cdot,\varphi^k\rangle\varphi^k \ , \quad ||\varphi^k|| = 1 \ .$$

Поскольку функция от оператора записывается как

$$f(\lambda) = \sum f(\lambda^k)\langle\cdot,\varphi^k\rangle\varphi^k \ ,$$

то резольвента в спектральном представлении оператора A имеет вид

$$R_\lambda(A) = \sum_k \frac{\langle\cdot,\varphi^k\rangle\varphi^k}{\lambda_k-\lambda} \ . \tag{21}$$

Подставляя в (21) вместо φ^k нормированные на единицу собственные функции оператора либо второй производной либо собственные векторы разностного оператора, а вместо λ_k соответствующие собственные значения $\frac{\pi^2k^2}{b^2}$ оператора двойного дифференцирования или собственные числа $\frac{2}{h^2}(1-\cos\frac{\pi k}{b}h)$ разностного, мы получим, соответственно, резольвенту оператора второй производной или разностной второй производной.

Пусть $\rho_n^2 = \int\limits_0^b \sin^2(\frac{\pi n x}{b})dx$, тогда нормированные собственные функции оператора двойного дифференцирования имеют вид $F^{(n)} = \frac{1}{\rho_n}\sin\frac{\pi n}{b}x$. В случае разностного оператора положив

$$\rho_n^2 = \sum_{j=1}^N \sin^2\frac{\pi n}{b}x_j \ ,$$

получаем нормированные собственные векторы $\mathbf{F}^n$ с компонентами

$$F_j^n = \frac{1}{\rho_n}\sin\frac{\pi n}{b}x_j = \frac{1}{\rho_n}\sin\frac{\pi n}{b}hj \ .$$

Получим матричные элементы резольвенты разностного оператора. В базисе из собственных векторов разностного оператора, резольвента, очевидно, представляется диагональной матрицей. Пусть $\mathbf{e}^1,\mathbf{e}^2,\ldots,\mathbf{e}^N$ — некоторый ортонормированный базис в $\mathbf{R}^N$ и $\mathbf{v}$ — произвольный вектор. Разложим $\mathbf{v}$ и собственные векторы $\mathbf{F}^n$ по этому базису

$$\mathbf{v} = \sum_{i=1}^N \langle\mathbf{v},\mathbf{e}^i\rangle\mathbf{e}^i = \sum_{i=1}^N v_i\mathbf{e}^i \ , \qquad \mathbf{F}^n = \sum_{i=1}^N \langle\mathbf{F}^n,\mathbf{e}^i\rangle\mathbf{e}^i = \sum_{i=1}^N F_i^n\mathbf{e}^i \ .$$

Действие резольвенты на $\mathbf{v}$ имеет вид

$$R_\lambda(A)\mathbf{v} = \sum_{i=1}^N \frac{\langle\mathbf{v},\mathbf{F}^i\rangle\mathbf{F}^i}{\lambda_i-\lambda} = \sum_{i=1}^N \frac{\sum\limits_{l=1}^N v_l F_l^i}{\lambda_i-\lambda}\mathbf{F}^i \ .$$

k-ая компонента вектора $R_\lambda(A)v$ есть

$$[R_\lambda(A)v]_k = \sum_{i=1}^N \frac{\sum\limits_{l=1}^N v_l F_l^i}{\lambda_i-\lambda}F_k^i = \sum_{i=1}^N\sum_{l=1}^N \frac{F_l^i F_k^i}{\lambda_i-\lambda}v_l \ ,$$

то есть матричные элементы оператора $R_\lambda(A)$ имеют вид:

$$R_\lambda(A)_{kl} = \sum_{i=1}^{N} \frac{F_l^i F_k^i}{\lambda_i - \lambda} \ .$$

Верхний индекс у F нумерует собственные функции, нижний индекс — их компоненты.

3.5.4 Теория возмущений

Спектр оператора двойного дифференцирования и спектр соответствующего разностного оператора нам известен. Рассмотрим соответствующие возмущенные задачи:

$$\begin{cases} -\frac{d^2}{dx^2}\Psi + \varepsilon q(x)\Psi = \lambda\Psi, \\ \Psi(0) = \Psi(b) = 0, \end{cases} \qquad \begin{cases} -\frac{F_{i-1} - 2F_i + F_{i+1}}{h^2} + \varepsilon q_i F_i = \lambda F_i \ , \\ F_0 = F_{N+1} = 0 \ . \end{cases}$$

Здесь ε — малый параметр, q — потенциал.

Изложим суть метода теории возмущений [18] для случая оператора с дискретным спектром. Пусть A и Q — два сомосопряженных оператора, причем собственные функции и собственные значения A известны:

$$A\Psi^k = \lambda_k \Psi^k \ .$$

Требуется провести приближенно спектральный анализ возмущенного оператора $A + \varepsilon Q$, то есть найти решения задачи

$$[A + \varepsilon Q]\varphi^k = \mu_k \varphi^k \ . \tag{22}$$

Будем предполагать, что спектр A невырожден. Разложим собственные значения и собственные функции возмущенного оператора по степеням малого параметра ε :

$$\mu_k = \lambda_k + \varepsilon\mu_k^{(1)} + \varepsilon^2\mu_k^{(2)} + \ldots \ , \tag{23}$$

$$\varphi^k = \Psi_k + \varepsilon\varphi_k^{(1)} + \varepsilon^2\varphi_k^{(2)} + \ldots \ , \tag{24}$$

где $\lambda_k^{(i)}$ и $\varphi_k^{(i)}$ — некоторые неизвестные числа и функции, соответственно. Ограничимся первым порядком теории возмущений. Подставляя в (22) выражения (23), (24) и учитывая само уравнение $A\Psi^k = \lambda_k \Psi^k$, с точностью до членов первого порядка по ε получаем

$$[A + \varepsilon Q - \lambda_k - \varepsilon\mu_k^{(1)}](\Psi^k + \varepsilon\varphi_k^{(1)}) = 0 \ ,$$

или

$$\underbrace{(A - \lambda_k)\Psi_k}_{0} + \varepsilon[(A - \lambda_k)\varphi_k^{(1)} + (Q - \mu_k^{(1)})\Psi_k] = 0 \ .$$

Таким образом, необходимо решить уравнение:

$$(A - \lambda_k I)\varphi_k^{(1)} = (\mu_k^{(1)} - Q)\Psi_k \ . \tag{25}$$

Обозначим $A - \lambda_k I = B$. Это вырожденный оператор (поскольку имеет нулевое собственное значение: $B\Psi^k = 0$). Пусть также $(\mu_k^{(1)} - Q)\Psi_k = g$. Тогда задача сводится к уравнению

$$B\varphi_k^{(1)} = g \ .$$

В соответствии с альтернативами Фредгольма, эта задача имеет единственное решение, если функция g ортогональна ядру сопряженного оператора, то есть решениям задачи $B^*g = 0$. Не вдаваясь в доказательства, поясним этот результат следующим образом. Представим g в виде суммы двух функций, одна из которых принадлежит ядру сопряженного оператора, а другая ортогональному дополнению: $g = v^1 + v^2$, $v^1 \perp v^2$, $B^*v^1 = 0$. Тогда

$$||g||^2 = \langle B\varphi_k^{(1)}, g\rangle = \langle \varphi_k^{(1)}, B^*(v^1 + v^2)\rangle = \langle B\varphi_k^{(1)}, v^2\rangle = \langle v^1 + v^2, v^2\rangle = \langle v^2, v^2\rangle ,$$

то есть норма не зависит от проекции g на ядро сопряженного оператора, иначе говоря, этой проекции просто нет. В нашей ситуации $B = A - \lambda_k I$ — самосопряженный оператор. Таким образом, условие разрешимости (25) принимает вид $(\mu_k^{(1)} - Q)\Psi_k \perp \Psi_k$ или $\langle(\mu_k^{(1)} - Q)\Psi_k, \Psi_k\rangle = 0$, откуда

$$\mu_k^{(1)} = \langle Q\Psi_k, \Psi_k\rangle .$$

Таким образом, поправки к собственным значениям определены. Поправки к собственным функциям определяем из того же уравнения (25)

$$(A - \lambda_k)\varphi_k^{(1)} = (\mu_k^{(1)} - Q)\Psi^k .$$

То есть формально

$$\varphi_k^{(1)} = R_{\lambda_k}(A)(\mu_k^{(1)} - Q)\Psi^k = \sum_{i=1} \frac{\langle \Psi^i, (\mu_k^{(1)} - Q)\Psi^k\rangle}{\lambda_i - \lambda_k}\Psi_k^i .$$

Но $R_\lambda(A)$ при $\lambda = \lambda_k$ не является ограниченным оператором. С другой стороны $\langle(\mu_k^{(1)} - Q)\Psi_k, \Psi_k\rangle = 0$, поэтому суммирование можно вести по $i \neq k$. Продолжая равенство, получаем

$$\varphi_k^{(1)} = \sum_{i \neq k} \frac{\langle \Psi^i, (\mu_k^{(1)} - Q)\Psi^k\rangle}{\lambda_i - \lambda_k}\Psi^i = \sum_{i \neq k} \frac{\langle \Psi^i, Q\Psi^k\rangle}{\lambda_k - \lambda_i}\Psi^i .$$

Здесь мы воспользовались тем, что собственные функции ортогональны. Итак, в первом порядке теории возмущений

$$\mu_k = \lambda_k + \varepsilon\langle Q\Psi_k, \Psi_k\rangle ,$$

$$\varphi_k = \Psi_k + \varepsilon \sum_{i \neq k} \frac{\langle \Psi^i, Q\Psi^k\rangle}{\lambda_k - \lambda_i}\Psi^i .$$

Литература

[1] *Н.Н. Калиткин* // Численные методы // Москва, Наука, 1978.

[2] *Н.С.Бахвалов, Н.П.Жидков, Г.М.Кобельков* // Численные методы // Москва — Санкт-Петербург, Лаборатория базовых знаний, 2000.

[3] *Д.Каханер, К.Моулер, С.Неш* // Численные методы и программное обеспечение // Москва, Мир, 1998.

[4] *Дж. Форсайт, М.Малькольм, К.Моулер* // Машинные методы математических вычислений // Москва, Мир, 1980.

[5] *С.Б. Стечкин, Ю.Н. Субботин* // Сплайны в вычислительной математике // Москва, Наука, 1976.

[6] *Дж.Бейкер, П.Грейвс-Моррис* // Аппроксимации Паде // Москва, Мир, 1986.

[7] *Д.Мак-Кракен, У.Дорн* // Численные методы и программирование на ФОРТРАНе // М., Мир, 1977.

[8] *В.В.Вершинин, Ю.С.Завьялов, Н.Н.Павлов* // Экстремальные свойства сплайнов и задача сглаживания // Новосибирск, Наука, 1988.

[9] *А.И.Гребенников* // Метод сплайнов и решение некорректных задач теории приближений // Издательство МГУ, 1983.

[10] *Э.Дулан, Дж.Миллер, У.Шилдерс* // Равномерные численные методы решения задач с пограничным слоем // М., Мир, 1983.

[11] *В.В.Воеводин, Ю.А.Кузнецов* // Матрицы и вычисления // М., Наука, 1984.

[12] *С.Писсанецки* // Технология разреженных матриц // М., Мир, 1988.

[13] *И.С.Березин, Н.П.Жидков* // Методы вычислений. Т.1. // М., Наука, 1966.

[14] *И.С.Березин, Н.П.Жидков* // Методы вычислений. Т.2. // М., Физматгиз, 1962.

[15] *А.Н.Колмогоров, С.И.Фомин* // Элементы теории функций и функционального анализа // М., Наука, 1972.

[16] *Д.К.Фаддеев*// Лекции по алгебре// М., Наука, 1984.

[17] *Г.Е.Шилов* // Математический анализ (функции одного переменного. Часть 3) // М., Наука, 1970.

[18] *Л.Д.Ландау, Е.М.Лифшиц* // Квантовая механика (нерелятивистская теория) // М., Наука, 1989.

[19] *А.Н.Тихонов, А.А.Самарский* // Уравнения математической физики // М., Наука, 1972.

[20] *Г.Корн, Т.Корн* // Справочник по математике // М., Наука, 1984.

Оглавление

1 Системы уравнений **3**

1.1 Решение нелинейных уравнений 3

1.1.1 Одномерный случай 3

1.1.2 Метод Ньютона 4

1.1.3 Метод секущих 5

1.1.4 Многомерный случай 7

1.2 Решение линейных систем 8

1.2.1 Обусловленность линейных систем, погрешность 8

1.2.2 Метод Гаусса 9

1.2.3 L-R разложение 11

1.2.4 Метод прогонки 12

1.2.5 Метод итераций для решения линейных систем 14

1.2.6 Метод Зейделя 15

2 Алгебраические спектральные задачи **18**

2.1 Некоторые сведения из матричной теории 18

2.2 Собственные числа эрмитовых матриц 19

2.2.1 Интерполяционный метод 19

2.2.2 Нахождение максимального по модулю собственного значения 19

2.2.3 Обратные итерации 21

2.3 Неэрмитовы матрицы 21

2.3.1 Дополнительные сведения 21

2.3.2 Метод итераций для максимального по модулю собственного числа кратности 2 в случае жордановой аномалии 22

3 Обыкновенные дифференциальные уравнения **25**

3.1 Общие сведения 25

3.1.1 Задача Коши 25

3.1.2 Краевая задача 27

3.1.3 Задача Штурма-Лиувилля 27

3.1.4 Что понимается под численным решением 27

3.2 Задача Коши 27

3.2.1 Получение явных схем 28
3.2.2 Схема Эйлера (метод ломаных) 28
3.2.3 Методы Рунге-Кутта 29
3.2.4 Методы Адамса 31
3.3 Краевая задача 32
3.3.1 Метод стрельбы 32
3.3.2 Метод сеток (разностный метод) 33
3.3.3 Сходимость сеточных методов 33
3.3.4 Метод Нумерова 34
3.4 Задача Штурма-Лиувилля 35
3.4.1 Метод стрельбы 35
3.4.2 Метод сеток 35
3.5 Разностный оператор второй производной 36
3.5.1 Оператор второй производной 36
3.5.2 Разностный оператор 36
3.5.3 Резольвента 38
3.5.4 Теория возмущений 39